

Olga Zlatkin-Troitschanskaia
Hans Anand Pant · Miriam Toepper
Corinna Lautenbach *Editors*

Student Learning in German Higher Education

Innovative Measurement Approaches
and Research Results



Springer VS

Student Learning in German Higher Education

Olga Zlatkin-Troitschanskaia ·
Hans Anand Pant · Miriam Toepper ·
Corinna Lautenbach
Editors

Student Learning in German Higher Education

Innovative Measurement
Approaches and Research Results

 Springer VS

Editors

Olga Zlatkin-Troitschanskaia
Johannes Gutenberg University Mainz
Mainz, Germany

Hans Anand Pant
Humboldt University of Berlin
Berlin, Germany

Miriam Toepfer
Johannes Gutenberg University Mainz
Mainz, Germany

Corinna Lautenbach
Humboldt University of Berlin
Berlin, Germany

ISBN 978-3-658-27885-4 ISBN 978-3-658-27886-1 (eBook)
<https://doi.org/10.1007/978-3-658-27886-1>

© Springer Fachmedien Wiesbaden GmbH, part of Springer Nature 2020, corrected publication 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer VS imprint is published by the registered company Springer Fachmedien Wiesbaden GmbH part of Springer Nature.

The registered company address is: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Contents

1 Modeling and Measuring Competencies in Higher Education – The KoKoHs Program	1
<i>Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., and Lautenbach, C.</i>	
2 Teachers’ Competence Development	7
2.1 Putting Educational Knowledge of Prospective Teachers to the Test – Further Development and Validation of the BilWiss Test	9
<i>Kunter, M., Kunina-Habenicht, O., Holzberger, D., Leutner, D., Maurer, C., Seidel, T., and Wolf, K.</i>	
2.2 Analyses and Validation of Central Assessment Instruments of the Research Program TEDS-M	29
<i>Kaiser, G., and König, J.</i>	
2.3 Planning Competence of Pre-Service German Language Teachers – Conceptualization, Measurement, and Validation	53
<i>König, J., Bremerich-Vos, A., Buchholtz, C., Fladung, I., and Glutsch, N.</i>	

2.4 Relationships between Domain-Specific Knowledge, Generic Attributes, and Instructional Skills – Results from a Comparative Study with Pre- and In-Service Teachers of Mathematics and Economics	75
<i>Kuhn, C., Zlatkin-Troitschanskaia, O., Lindmeier, A., Jeschke, C., Saas, H., and Heinze, A.</i>	
2.5 Development of Prospective Physics Teachers’ Professional Knowledge and Skills during a One-Semester School Internship	105
<i>Vogelsang, C., Borowski, A., Kugelmeyer, C., Riese, J., Buschhüter, D., Enkrott, P., Kempin, M., Reinhold, P., Schecker, H., and Schröder, J.</i>	
2.6 Linguistically Responsive Teaching in Multilingual Classrooms – Development of a Performance-Oriented Test to Assess Teachers’ Competence	125
<i>Lemmrich, S., Hecker, S.-L., Klein, S., Ehmke, T., Koch-Priewe, B., Köker, A., and Ohm, U.</i>	
2.7 Effects of Early Childhood Teachers’ Mathematics Anxiety on the Development of Childrens’ Mathematical Competencies	141
<i>Jenßen, L., Hosoya, G., Jegodtka, A., Eilerts, K., Eid, M., and Blömeke, S.</i>	
3 Generic Competencies and Their Impact on Learning in Higher Education	163
3.1 Modelling, Assessing, and Promoting Competences for Self-Regulated Learning in Higher Education	165
<i>Eckerlein, N., Dresel, M., Steuer, G., Foerst, N., Ziegler, A., Schmitz, B., Spiel, C., and Schober, B.</i>	

3.2 The Relationship between General Intelligence and Media Use among University Students	181
<i>Jitomirski, J., Zlatkin-Troitschanskaia, O., and Schipolowski, S.</i>	
3.3 Multiple Document Comprehension of University Students – Test Development and Relations to Person and Process Characteristics	221
<i>Schoor, C., Hahnel, C., Mahlow, N., Klagges, J., Kroehne, U., Goldhammer, F., and Artelt, C.</i>	
3.4 What Does It Take to Deal with Academic Literature? Epistemic Components of Scientific Literacy	241
<i>Münchow, H., Richter, T., and Schmid, S.</i>	
3.5 Measuring Scientific Reasoning Competencies – Multiple Aspects of Validity	261
<i>Krüger, D., Hartmann, S., Nordmeier, V., and Upmeyer zu Belzen, A.</i>	
3.6 Performance Assessment of Generic and Domain-Specific Skills in Higher Education Economics	281
<i>Nagel, M.-T., Zlatkin-Troitschanskaia, O., Schmidt, S., and Beck, K.</i>	
3.7 The Research Group Performance-Based Assessment of Communication in KoKoHs – a Bridge between Educational Theory and Empirical Educational Research	301
<i>Falkenstern, A., Schwabe, U., Walz, K., and Braun, E.</i>	

4 Domain-Specific Competencies in Business, Economics and Medicine	315
4.1 Measuring Medical Competence and Entrusting Professional Activities in an Assessment Simulating the First Day of Residency	317
<i>Prediger, S., Berberat, P. O., Kadmon, M., and Harendza, S.</i>	
4.2 Impact of Affective-Motivational Dispositions on Competence in Sustainability Management	333
<i>Michaelis, C., Aichele, C., Hartig, J., Seeber, S., Dierkes, S., Schumann, M., Jan Moritz, A., Siepelmeyer, D., and Repp, A.</i>	
4.3 The Impact of Entry Preconditions on Student Dropout and Subject Change in Business and Economics	351
<i>Kühling-Thees, C., Happ, R., Zlatkin-Troitschanskaia, O., and Pant, H. A.</i>	
4.4 Influences on the Development of Economic Knowledge over the First Academic Year – Results of a Germany-Wide Longitudinal Study	371
<i>Schlx, J., Zlatkin-Troitschanskaia, O., Kühling-Thees, C., and Brückner, S.</i>	
4.5 Influences on Master’s Degree Students’ Economic Knowledge	401
<i>Kraitzek, A., Förster, M., and Zlatkin-Troitschanskaia, O.</i>	
Correction to: Measuring Scientific Reasoning Competencies – Multiple Aspects of Validity	261
<i>Krüger, D., Hartmann, S., Nordmeier, V., and Upmeyer zu Belzen, A.</i>	



Modeling and Measuring Competencies in Higher Education

1

The KoKoHs Program

Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., and Lautenbach, C.

Over the past decade, tertiary education has increasingly been gaining importance in society. Developments such as the continuously growing number of students in higher education and increasing student mobility have raised questions of efficiency and effectiveness in tertiary education, calling for valid assessments of competencies and student learning outcomes. Assessments of the output of higher education can yield important evidence regarding the effectiveness of this highly important educational sector and thus provide a basis for improvement measures at the individual and institutional levels (Coates and Zlatkin-Troitschanskaia 2019).

Modern higher education focuses on the acquisition of domain-specific knowledge and on the development and promotion of generic (interdisciplinary) skills (e.g. critical thinking), which, according to current surveys amongst employers, are increasingly gaining significance in the 21st century (Association of German Chambers of Industry and Commerce (DIHK) 2015). Such a competence portfolio, acquired over the course of academic studies, is crucial for all professionals and globally engaged citizens and allows for lifelong learning, which is necessary in today's continuously changing age of information.

Despite this sociopolitical consensus and the growing competence orientation (in the context of the Bologna reform), there have been only few evidence-based in-

sights into this field up until the last decade, particularly regarding the competencies of higher education students. Therefore, the German Federal Ministry of Education and Research established the Germany-wide research initiative “Modeling and Measuring Competencies in Higher Education (KoKoHs)”¹ in 2011 and – after a positive external evaluation in 2015 – decided to continue to fund this research in the context of the German program “Modeling and Measuring Competencies in Higher Education (KoKoHs) – Validation and Methodological Innovations” until 2020.

In the first research program, KoKoHs (2011–2015), more than 220 researchers from various fields such as subject-specific didactics, learning psychology, and psychometrics developed first modeling approaches and the corresponding measuring instruments for the valid assessment of student competencies in the context of 24 collaborative research projects at over 70 universities and research institutes, focused on central study domains such as business and economics, engineering, and teacher education (for a detailed description of the first KoKoHs research program (2011–2015) and the individual projects and results, see Zlatkin-Troitschanskaia et al. 2017). These models and tools developed in KoKoHs were one of the key results of this first working phase, which ran until 2015. Another equally important outcome of this research phase were the findings on students’ competence levels in different study phases, which revealed many deficits. At the same time, the generalizability of these results was questionable, as some of the newly developed KoKoHs instruments had not yet been comprehensively validated in accordance with a number of validation criteria as recommended in the Standards for Psychological and Educational Testing by AERA et al. (2014). Another shortcoming had been the fact that most of the newly developed test instruments were paper-pencil-based and altogether only few innovative assessments had been developed in the first phase. Based on the results and recommendations from an international audit at the end of the first phase, the second phase of the KoKoHs program was launched nationwide in 2015 with a focus on validation and methodological evaluations.

In this follow-up research program, KoKoHs (2016–2020), more than 100 researchers comprehensively validated KoKoHs assessments and developed new innovative modeling approaches and the corresponding measuring instruments for the valid assessment of student competencies in the context of 16 collaborative research projects at over 40 universities and research institutes, again focused on central study domains such as business and economics as well as teacher education. In this program, one new study domain was included: medicine. Moreover, some of

1 For further information on KoKoHs, see <https://www.blogs.uni-mainz.de/fb03-kokohs-eng/>

the projects focused on transferring and adapting modeling approaches and assessments from one domain to another (e.g. from mathematics to economics). Overall, this program consists of three large clusters: four projects focusing on domain-specific competencies in economics and medicine, five projects with a focus on domain-independent competencies such as scientific reasoning and self-regulation skills (for domain-specific and generic competencies, see Zlatkin-Troitschanskaia, Pant, and Greiff 2019), and the largest cluster with seven projects and a focus on teacher education in different domains such as mathematics, physics, or economics (for teachers' competencies, see Cortina, Pant, and Zlatkin-Troitschanskaia 2019).

A common focus of all projects was the in-depth validation of KoKoHs assessments following the validation criteria of AERA et al. (2014). Most projects were also characterized by their focus on the development and validation of complex technology-based assessments, which are mostly performance-oriented (for performance assessment, see Zlatkin-Troitschanskaia and Shavelson 2019). Innovative technology-based test formats such as computer-based learning diaries or mobile apps were also developed and implemented. In this research phase, some of the projects have had a longitudinal design, which has allowed for valid statements about the development of competencies over the course of academic studies. In addition, several instruments developed and validated in KoKoHs have now also been tested and used in many other countries such as Japan, the US, and China, and comparative analyses have already been carried out (for cross-national studies, see Zlatkin-Troitschanskaia et al. 2018).

Overall, in the 40 collaborative KoKoHs projects (which, in turn, comprised about 100 individual projects), theoretical-conceptual competence models and corresponding measurement instruments were developed and successfully validated for selected large study domains (e.g. economics, teacher education, STEM). These models differentiate, reliably describe, and assess the competences of students in different phases of higher education – entry, undergraduate and postgraduate studies. Over 100 newly developed innovative video-, computer-, and simulation-based test instruments were validated across Germany at more than 350 universities with over 75,000 undergraduate and master's students. The assessments focused on both discipline-specific competencies and generic skills, which students and graduates should acquire over the course of their studies and which employers and other stakeholders expect according to the professional and social requirements of the 21st century.

Building on best practices from the first funding phase of the KoKoHs program (2011–2015), the subsequent funding phase ran from 2015 to 2020 and brought together experts from various fields and with different methodological backgrounds in cross-university project alliances within a joint international and interdiscipli-

nary research network. Based on the models and instruments for the reliable and valid assessment of competencies acquired in various study domains in higher education that were developed and empirically tested in the first funding phase, this follow-up research phase of KoKoHs aimed to increase the explanatory power and broaden the scope of use of the KoKoHs test instruments through in-depth validation and to drive methodological innovation in higher education competency assessment.

KoKoHs is the only existing nationwide program in which students' learning outcomes in higher education are systematically, validly, and objectively assessed and analyzed. The KoKoHs program provides unique findings on the acquisition and development of students' competencies in German higher education, which form a significant basis for the optimization of learning and teaching practice.

This book is based on the research and development work conducted in KoKoHs over the past decade and offers a comprehensive overview of current innovative tools and approaches to assessing domain-specific and generic student learning processes and learning outcomes in higher education. It presents the work of all KoKoHs projects, thus offering an insight into the most significant research program focused on student learning outcomes in higher education to date. In this volume, innovative modeling and measuring approaches as well as the newly developed objective, valid, and reliable assessment tools for student learning in higher education are presented and critically discussed, with a particular focus on using the developed models and assessments in both further research and higher education practice.

In addition to presenting key conceptual and methodological findings from work within the KoKoHs program, the 88 authors in this book also present key research results and lessons learned from their research to provide new insights into how student learning in higher education can be assessed in various contexts and to show what we can learn from the assessment results. Most contributions also provide an outlook on possible approaches to implementing the instruments into teaching and learning practice and transfer studies. The authors also give a few examples of how higher education practitioners in particular can effectively support teaching and learning at their universities by using the KoKoHs assessments and tools.

With its very broad spectrum of contributions focused on both innovative research and the practical application of assessments in higher education, this volume offers valuable insights for scientists in higher education research as well as related disciplines such as psychology, educational sciences, lecturers in university practice, university evaluation, accreditation agencies, higher education pol-

icy-makers, students, companies and all other stakeholders interested in higher education student learning outcomes.

We would like to thank everyone who contributed to this book. This includes, of course, the 88 authors from the KoKoHs projects and all of the researchers and student assistants who contributed to the work conducted in the KoKoHs program and documented in this volume. We would like to thank all national and international critical advisors of this program, especially *Daniel Koretz*, *Fritz Oser*, *James Pellegrino*, and *Richard Shavelson*, who have significantly supported the work conducted in this program over the past decade. Our sincere thanks also go to all of our colleagues who provided external reviews of the contributions and thus contributed significantly to the quality of the articles in this volume. Special thanks go to the sponsor of the KoKoHs program, the German Federal Ministry for Education and Research, which, thanks to its long-term support, has enabled us to carry out sustainable research and development in this field for almost a decade now, thus also contributing to the emergence of a new field of research and to establishing empirical research in higher education in a sustainable manner. In this context, we would like to thank *Martina Diegelmann* in particular, who has critically supervised the program over the past decade and has decisively contributed to its structural and conceptual development. We would also like to thank the DLR project management agency for providing administrative support to all KoKoHs projects.

Many others were involved in the preparation of this book, including our student assistants in KoKoHs and *Mirco Kunz* in particular, who was responsible for the technical preparation of the manuscript, as well as our staff members from the field of translation studies, *Katja Kirmizakis* and *Annika Weibell*, who proofread the contributions in this volume as well as this article.

References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Association of German Chambers of Industry and Commerce (DIHK) (2015). *Licht und Schatten (DIHK-Onlineumfrage zur Berufsschulsituation in den IHK-Regionen)*. Verfügbar unter <http://www.dihk.de/themenfelder/aus-und-weiterbildung/schule-hochschule/schule/umfragen-und-prognosen/dihk-berufsschulumfrage>
- Coates, H., & Zlatkin-Troitschanskaia, O. (2019). The Governance, Policy and Strategy of Learning Outcomes Assessment. *Higher Education Policy*, 32(19), 1–6.

- Cortina, K., Pant, H. A., & Zlatkin-Troitschanskaia, O. (2019). Kompetenzerwerb zukünftiger LehrerInnen in der universitären Ausbildung. [Themenheft]. *Zeitschrift für Pädagogik* (4/2019).
- Zlatkin-Troitschanskaia, O., Pant, H. A., & Greiff, S. (2019). Assessing Academic Competencies in Higher Education. [Special Issue]. *Zeitschrift für Pädagogische Psychologie*, 33(2).
- Zlatkin-Troitschanskaia, O., & Shavelson, R. J. (2019). Performance Assessment of Student Learning in Higher Education [Special issue]. *British Journal of Educational Psychology*, 89(3).
- Zlatkin-Troitschanskaia, O., Pant, H. A., Lautenbach, C., Molerov, D., Toepper, M., & Brückner, S. (2017). *Modeling and Measuring Competencies in Higher Education. Approaches to Challenges in Higher Education Policy and Practice*. Wiesbaden: Springer.
- Zlatkin-Troitschanskaia, O., Toepper, M., Pant, H. A., Lautenbach, C., & Kuhn, C. (Eds.) (2018). *Assessment of Learning Outcomes in Higher Education – Cross-national Comparisons and Perspectives*. Wiesbaden: Springer.



2.1 Putting Educational Knowledge of Prospective Teachers to the Test

Further Development and Validation of
the BilWiss Test

Kunter, M., Kunina-Habenicht, O.¹, Holzberger, D., Leutner, D.,
Maurer, C., Seidel, T., and Wolf, K.

Abstract

Teachers' generic educational knowledge theoretically constitutes an aspect of their professional competence. However, empirical evidence for its importance for teachers' daily practice is scarce. In this chapter, we describe findings from the BilWiss research program, which aimed to investigate the development and relevance of the type of generic educational knowledge typically addressed in university teacher education. We developed a standardized test that assesses generic knowledge in the following six domains: learning and development, instruction, assessment, educational theory (and history), school system and educational policy, and the teaching profession. We present findings from a series of studies that (a) provide evidence for the validity of the test score interpretations and (b) prove the predictive value of this test for diverse professional activities. These results are discussed regarding their theoretical and practical implications for teacher education.

1 Mareike Kunter and Olga Kunina-Habenicht are both first authors of this contribution.

Keywords

Educational knowledge; pedagogical knowledge, assessment, knowledge test; validity, teacher education, test development

1 Teachers' Educational Knowledge as an Aspect of Their Professional Competence

In research on teachers' professional competence there is a consensus that teachers' success in providing high-quality instruction is connected to their profession-specific declarative and procedural knowledge (Baumert et al. 2010; Kennedy et al. 2008; Schleicher 2016; Shulman 1986). Thus, the foundations of professional competence are partly laid in the theoretical part of teacher education at university, which aims particularly at providing subject-specific and generic knowledge in formal learning settings. While the importance of a professional knowledge base per se is uncontested, it is less clear what content such a knowledge base should include (Zeichner 2005).

In his seminal work, Shulman (1986) distinguishes between content knowledge, pedagogical content knowledge, and pedagogical knowledge. The first two describe subject-specific knowledge, which has received much research attention in recent years. There is ample evidence that knowledge about the subject matter itself, subject-specific forms of instruction, and typical student thinking in a domain are important prerequisites for high-quality instruction and, thus, student learning (Abell 2008; Baumert et al. 2010; Depaepe et al. 2013; Hill et al. 2005). The part of teacher knowledge that transcends subject matter has received much less attention in current research. Shulman (1986) defines pedagogical knowledge as "knowledge of generic principles of classroom organization and management" and notes that "proper professional board examination would include other equally important sections as <...> knowledge of general pedagogy, knowledge of learners and their backgrounds, principles of school organization, finance and management, and the historical, social, and cultural foundations of education" (p. 14). In line with Shulman's argumentation we use the term "educational knowledge" that extends the narrow conception of pedagogical knowledge and define it as "teachers' subject-unspecific professional knowledge that comprises both classroom-related topics (instruction, learning and development, and assessment) as well as context-related topics (e.g., knowledge on the educational system, school development, or educational theory and history)" (Linninger et al. 2015, p. 73).

Voss et al. (2015) provided a comprehensive overview of recent developments and assessments measuring pedagogical teacher knowledge. This review revealed that previous research mainly focused on pedagogical topics that are closely related to instruction (e.g., classroom management, learning support) and largely ignored topics addressing matters outside the classroom such as principles of school organization, historical foundations of education, or knowledge about the teacher profession. Previous research revealed small significant correlations between pedagogical knowledge (in a narrow sense) and the instructional quality rated by school students (König and Pfanzl 2016; Voss et al. 2014). Nevertheless, it is not clear which benefit the broader generic educational knowledge might have both for teaching situations and for situations outside of classroom.

To close this gap, the BilWiss² research program aimed to investigate both the empirical structure of educational knowledge and its determinants and consequences. In this chapter, we summarize the theoretical background and the main results of the BilWiss project.

2 Background: Educational Foundation Courses in Teacher Education

The structure of teacher education varies across and within countries. University teacher training usually involves courses covering general educational topics and the study of one or more specific subjects. In general, two different models of teacher education are common: the concurrent model where subject courses, educational courses, and practical experiences are combined within one course of study and the consecutive model where a disciplinary degree (i.e., subject-specific) is followed by a degree in education (EURYDICE, 2002). The relative importance of subject-specific and generic parts (expressed in allocated credit points) differs both between and within different countries and institutions (Schmidt et al. 2011; 2008). However, in all systems, teacher students have to attend courses that aim at providing the generic educational knowledge seen as particular to the profession of teachers, the so-called “educational foundation courses”.

2 BilWiss stands for „Bildungswissenschaftliches Wissen“, which is the German translation of “educational knowledge”. The full name of the research program is “Bildungswissenschaftliches Wissen als Teil professioneller Kompetenz in der Lehramtsausbildung [Educational Knowledge as a Part of Professional Competence in Teacher Education]”. For more information see: <https://bilwiss.paedpsych.de/>

Educational foundation courses have a long tradition as a relevant part of teacher education (for a historical review, see Tozer and McAninch 1986) and are assumed to provide important learning opportunities for the acquisition and construction of an educational knowledge base. Educational foundations are defined as a “broadly-conceived field of educational study that derives its <...> methods from a number of academic disciplines <...> including: history, philosophy, sociology, <...> psychology, <...>, educational studies <...>” (Council for Social Foundations of Education (CSFE) 1996, p. 3). In the last decades, there has been a vivid discussion about the nature of knowledge that should be taught within educational foundation courses in university teacher education programs (Hollins 2011; Patrick et al. 2011). Wilson and colleagues have reviewed studies on the impact of “pedagogical knowledge” on teacher effectiveness (Wilson et al. 2001; 2002) and conclude that “the impact of pedagogical knowledge or preparation was spotty and inconclusive” (2003, p. 16). Furthermore, the number of courses offered varies across institutions within single countries in terms of course sequencing and course content (Wilson et al. 2001, p. 12). Given the miscellany of topics from various disciplines (e.g., psychology, educational studies, sociology), it might be difficult for students to develop educational pedagogical knowledge in a sense of a coherent theoretical construct. The perceived fragmentation of educational courses occurs partly due to the high degree of freedom in the choice of educational courses (Terhart et al. 2010).

Furthermore, both teacher students and in-service teachers have often criticized university education and particularly, the educational foundation part, for providing insufficient practical preparation (Cochran-Smith and Zeichner 2005; Darling-Hammond et al. 2002; Veenman 1984; Zeichner 2006) and for the “absence of a set of organizing themes, shared standards, and clear goals”(Hollins 2011, p. 395). In contrast to this argumentation, we argue that – given the high complexity of teachers’ actions in rapidly changing situations with high demands on reflection ability (Leinhardt and Greeno 1986) – a thorough conceptual understanding of the domain and of educational topics is a key to improving teachers’ professional mastery of their job (Hollins 2011; Rittle-Johnson et al. 2001). To test this hypothesis, we refer to a theoretical framework which can serve as a foundation for the development of a research instrument that allows for objective and reliable assessment of educational knowledge. In line with this argumentation, there is empirical evidence that, despite the general criticism toward educational foundations, students and beginning teachers perceive these parts of university education as relevant for practical work (Alles et al. 2018; Dawson et al. 1984; Grossman and Richert 1988; Rösler et al. 2013).

In sum, although educational knowledge is deemed an important aspect of teachers' professional competence there is little consensus about how this knowledge could best be fostered during teacher education. One reason for this is insufficient empirical evidence on the structure of educational knowledge and its development during teacher education which is mainly due to the fact that to date, most studies have been based on self-report measures only (Choy et al. 2013; Wong et al. 2008). Very few researchers have attempted to measure generic educational knowledge directly via standardized assessments (Guerriero 2017; Sonmark et al. 2017; see also contributions in this volume).

The BilWiss research project started in 2009 to investigate the empirical structure, development, and impact of beginning teachers' educational knowledge. It was one of the first studies to develop a standardized knowledge test covering relevant content of educational foundation courses in academic teacher education. In the next section, we give an overview of the BilWiss research program including important findings.

3 The BilWiss Research Program

The BilWiss research program started in 2009 as a cooperation project between the Max Planck Institute for Human Development (Principal investigator: Jürgen Baumert), the Goethe University Frankfurt (Mareike Kunter), the University of Duisburg-Essen (Detlev Leutner), and the University of Münster (Ewald Terhart), joined in 2012 by the Technical University of Munich (Tina Seidel). The program, consisting of three consecutive funding periods, was supported by the German Federal Ministry of Education and Research (BMBF) over the course of ten years ending in spring 2019. The project team united of researchers who were all involved³ with teacher education and combined expertise in educational sciences and psychology, the main disciplines in educational foundation courses. In 2012, the project was complemented by an additional study funded by the Ministry of Education in the German state of North-Rhine Westphalia where the study took place. The aim of this study was the evaluation of a newly implemented induction program in this state.

3 Post doc and pre doc researchers involved in the BilWiss project: Andreas Dick, Theresa Dicke, Nora Hein, Olga Kunina-Habenicht, Hendrik Lohse-Bossenz, Christina Maurer, Nadine Schlomske-Bodenstein, Maria Schmidt, Franziska Schulze-Stocker, Kathleen Stürmer, Ziwen Teuber, Katharina Willis, Kristin Wolf.

The BilWiss research program addressed several research questions concerning the nature and meaning of teachers' educational knowledge. Empirically, these research questions were addressed with diverse methodological approaches as outlined below.

1) *Conceptualization of educational knowledge*: What is the central subject-unspecific content that prospective teachers should know at the end of their university studies? What are the central topics that educational foundational courses should address? What are important areas of knowledge and how can this knowledge be theoretically structured?

The empirical base for answering these research questions was literature work, curriculum analyses, and an expert Delphi study (Section 4.1).

2) *Assessment of educational knowledge*: Can educational knowledge be measured with a standardized knowledge test and can this test be used to describe knowledge differences between prospective teachers?

As outlined in Section 4.2, we constructed a comprehensive knowledge test that tapped all the topics identified as important by our expert study. After a series of pilot studies this test was administered to a representative sample of German teacher candidates after university completion and was later used in two longitudinal studies. In addition, various smaller validation studies were carried out.

3) *Relevance of educational knowledge*: To what degree does teachers' educational knowledge influence their later practice? What is the relationship between educational knowledge and teachers' instructional quality, their professional vision, other aspects of professional behavior, and teachers' professional well-being?

To investigate the practical relevance and long-term effect of teachers' educational knowledge on the successful mastery of their job, our research program included a longitudinal study that followed a sample of teacher candidates from completion of their university studies to their entrance into teaching practice and up to seven years beyond. To assess teachers' professional behavior, we used and developed various instruments such as video-based professional vision assessment, student ratings of instruction, vignette tests, and behavioral checklists (for results, see Section 4).

4) *Fostering educational knowledge*: How does educational knowledge develop during teacher education and beyond? How effective are educational foundation courses in fostering educational knowledge in students? Can tailored interventions support the growth of educational knowledge?

In addition to our first longitudinal study that investigated prospective teachers' development after university completion, our second longitudinal study targeted teacher students at university and investigated how their educational knowledge developed during the course of their university studies (covering a period of two years).

The next section provides a short overview of our core findings. Following the topics of the present volume, we then summarize the findings of the third and final research phase which focused on revision and validation of the BilWiss test.

4 Summary of Important Results from the BilWiss Research Program

4.1 Conceptualization of Educational Knowledge

Before the BilWiss project, the educational foundation courses in German teacher education were a matter of great debate. To tackle the much-debated heterogeneity of content and lack of consistence especially in this part of teacher education, the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (KMK) had passed their “Standards for Teacher Training in the Educational Sciences”. These standards specify the abilities and skills prospective teachers should acquire in the course of teacher education, specifically in the educational foundation courses (Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) 2014). However, curriculum analyses carried out in BilWiss (Schulze-Stocker 2016; 2017) and by others (Bauer et al. 2012; Hohenstein et al. 2014) showed that even several years later there was still much heterogeneity in terms of the content addressed in educational foundation courses. Further, as seen in these studies, many topics outlined in the Standards were not covered at all, and that universities varied substantially in the courses offered.

In a quest to identify an “ideal curriculum” that would include all the educational topics deemed important for prospective teachers, we carried out an expert survey using a Delphi-technique (Linstone and Turoff 1975). Forty-nine teacher education experts from heterogeneous fields (psychology, educational science, sociology, educators in teacher professional development) participated in a paper-pencil Delphi study in which they rated the importance of 213 potentially relevant educational topics (identified through curriculum analyses and from literature) in three consecutive rounds. These topics covered nine theoretically pre-defined content areas: instruction, educational theory, educational system, teacher profession, developmental processes, socialization processes, learning processes, dealing with diversity, and assessment. In spite of the heterogeneity found in educational foundation courses offered at university, experts substantially agreed which topics would be the most important for future teachers, leading to a list of 104 topics that were chosen by the majority of experts (for more detailed informa-

tion on specific instructions and consensus development within the Delphi Study, see Kunina-Habenicht et al. 2012; Lohse-Bossenz et al. 2013).

4.2 Assessment of Educational Knowledge: The Development of the BilWiss Test

The Delphi study provided us with a theoretical systematization of the generic educational knowledge that experts agreed should be imparted by university studies. This systematization was the foundation for our test construction. For each of the core Delphi topics we constructed at least one item. The constructed items capture either declarative or conceptual knowledge, with declarative items requiring the mere recall of certain facts or theories, and conceptual items requiring students to connect several theories or to apply certain theories to case examples. Experts from the disciplines of educational science and psychology were involved into the item construction process.

We conducted three pilot studies during the process of test development between June 2010 and November 2010. We optimized or excluded several items based on the examination of item difficulties, item discriminations, and frequencies of incorrect answer alternatives. The pilot studies showed that six dimensions of educational knowledge could be reliably distinguished: classroom teaching, learning and development (subsuming developmental, learning, and socialisation processes), creating school environments, theoretical educational foundations, assessment and evaluation, and teaching as a profession. The final test version consisted of 289 items and was administered in spring 2011 to 3298 persons in one federal German state (North-Rhine Westphalia), representing 87 per cent of the full cohort of recently-graduated teacher candidates at the beginning of their induction phase. To cover the whole breadth of educational knowledge within a reasonable test time, we used a multiple-matric booklet design with anchoring items and estimated person scores for the six dimensions using unidimensional 2-PL partial-credit IRT models. The reliabilities for all six scales were moderate or satisfactory, and the measurement models were replicated in a second independent sample (for further details, see Linninger et al. 2015).

In addition to the comprehensive long test version that distinguishes between six dimensions of educational knowledge, we created a short test version, which includes 57 items and provides a general score for educational knowledge (for an overview of the two test versions with item contents, reliabilities, and examples, see Linninger et al. 2015).

4.3 Relevance of Educational Knowledge: Does It Matter to Teacher Practice?

One of the main guidelines of the BilWiss project was the assumption that a good theoretical knowledge base is a necessary prerequisite for the professional behavior of teachers. The most important arena for teachers is undoubtedly their classroom and the classroom instruction. There is already some evidence that the quality of instruction and students' learning success may be influenced by teachers' pedagogical knowledge which we expected to confirm in our studies. However, following the broad conception of educational knowledge, we were also interested in teachers' professional activities apart from teaching, such as assessment, counselling, and engagement in school development.

Overall, our findings highlight the relevance of educational knowledge to teachers' practice, although our findings were not as clear-cut as expected. Regarding our assumption that educational knowledge works as a theoretical frame that allows for a functional analysis of professional situations, we found that teachers who scored high in our BilWiss test showed significantly more productive reflection when watching teaching scenes by other teachers than those who scored lower (Linninger et al. 2016). However, with regard to professional vision as assessed with the Observer Research Tool (Stürmer and Seidel 2015), no systematic relationships between the overall BilWiss test score and professional vision skills were found. This finding might be attributed to the fact that the Observer Research Tool is quite a focused tool, measuring professional vision skills in the context of three generic teaching and learning principles: clarity of learning goals, teacher support, and learning climate (Seidel et al. 2017). This strong focus might have led to a situation in which the two measurements are not validly linked to find systematic relationships.

Regarding our assumption that educational knowledge should also directly impact on teachers' behavior, we did find that teachers with greater educational knowledge reported greater improvement in teaching quality during their induction phase (Lohse-Bossenz et al. 2015); however, a direct link between teacher knowledge and instructional quality as perceived by students of their classes could not be established.

Going beyond classroom instruction, we also investigated the relevance of educational knowledge for other fields of teachers' work. An important finding was that educational knowledge, as measured by the BilWiss test (especially knowledge about classroom management or learning and development) works as a buffer against stress during the induction phase (Dicke et al. 2015a, b). Moreover, we found that teachers who scored higher in our subscale on knowledge about the

school system and educational policy at the end of the induction phase were more engaged in school development activities two years later than their less knowledgeable peers (Linninger et al. 2015). A pilot study has also shown that theoretical knowledge about counseling as measured within our BilWiss subscale “Teaching as a profession” was positively related to teachers’ projected counseling behavior in a situational judgement test (Maurer et al. 2018), a finding which we seek to replicate in ongoing analyses of our longitudinal sample.

4.4 Fostering Educational Knowledge

We found strong evidence that the educational knowledge we assessed in our BilWiss test represents professional academic knowledge actually acquired during teacher education and can be empirically distinguished from everyday notions on education that laypersons may hold. A qualitative study where we conducted cognitive interviews with persons during test taking revealed that teacher education graduates and advanced teacher students were often familiar with item topics from their studies and that they solved items mostly by retrieving academic knowledge gained in teacher education (Linninger et al. 2015). A comparison of the test scores of teacher education graduates with scores of first-semester teacher students or persons without teacher education showed substantial advantages for those who had completed teacher education (Kunina-Habenicht et al. 2013; Linninger et al. 2015). Moreover, we found higher test scores in different dimensions for teacher education graduates who had taken more courses in the respective domain (Schulze-Stocker et al. 2016).

5 Further Development and Validation: The BilWiss 2.0 Test

The first two phases of the BilWiss research program (2009–2016) showed that it is possible to measure prospective teachers’ educational knowledge via a standardized test, that this knowledge is actually a product of university teacher education, that beginning teachers differ substantially in their knowledge, and that these knowledge differences manifest in different qualities of behaviors. However, a number of issues remained open. First, although there was evidence of sufficient reliability and validity of the BilWiss test overall, the reliability of some scales was not satisfactory. In addition, item formats and the number of answer options in multiple-choice questions varied across the subscales of the test, which limited the

comparability of items and the scales, so that a further refinement and unification seemed warranted.

A second issue of concern was the lack of predictive validity of the subscale “instruction”, which did not show any associations with teaching-related outcome measures. A closer content inspection revealed that some of items in this scale – which all were selected mainly based on empirical grounds – did not show optimal match with the ranking of topics from the Delphi study, so that it seemed necessary to revise the instruction scale substantially.

A third issue was the limited economy of the test. On the one hand, the breadth of educational topics typically addressed in educational foundation courses required a large number of items to adequately cover the heterogeneity of the construct. On the other hand, we were aware of the limited applicability for a test with almost three hundred items and sought to optimize our short test which at that stage did not adequately cover the topics judged most important in our Delphi study.

A fourth critical issue concerned the generalizability of our test and results: As our prior work had been carried out just in one federal state in Germany, it was not clear to what degree our instrument and findings were applicable to other teacher education contexts. Finally, apart from measurement issues, we wanted to learn more about the changes in educational knowledge during the teacher education studies at the university.

The third phase of the BilWiss research program was thus dedicated to further development and validation of the BilWiss test.

5.1 Revision of the Original Test

In the first step, we identified dodgy items using prior data in terms of psychometric indicators and conducted cognitive labs using very similar procedures as described by Linninger and colleagues (2015) to identify items with high task-unspecific variance (e.g., items that could be solved by guessing or ambiguous wording). We rephrased the identified items and created some new tasks. In particular, we substantially modified the instruction scale, which includes many completely new items in the revised version. Moreover, we unified the number of answer options in multiple-choice questions to four in the entire test and ensured that the highly rated topics from the Delphi study were covered by at least one item in the short test form. The modified and new items were tested in an iterative process in several field tests (Kunina-Habenicht et al. 2020).

The short version of the revised BilWiss-2.0 test includes 65 items, the long version contains 119 items, all from the following six knowledge domains: learning and development, instruction, assessment, educational theory (and history), school system and educational policy, and teacher profession. 2-PL partial-credit IRT models were applied to a data set collected from 788 teacher students from four different German universities in four different states (for details, see Kunina-Habenicht et al. 2020).

5.2 Content Validity

All items of the test were examined in a qualitative study where 40 teacher students worked through the test (on average five teacher students per item) and verbalized their solution approaches, difficulties, and sources of knowledge. These cognitive labs showed that the majority of items were congruent with the topics addressed in educational foundation courses at university and that they could be answered by drawing on the content learned during university teacher education. The interviews also gave hints to construct-irrelevant variance inherent in some items that were considered during the item revision described above.

In two additional validation studies, we investigated whether the test content corresponded to the intended academic teacher education curriculum at the federal level in Germany (Kunina-Habenicht et al. 2019). In the first study delegates from the Ministries of Education of most German federal states (besides North-Rhine Westfalia) rated the relevance of the Delphi topics for their specific state. In the second study the test content (i.e. individual test items) was matched to the federal standards on the generic, educational part of academic teacher education (KMK, 2014). Results from both studies indicated that the BilWiss-2.0-test can be used across German federal states (Kunina-Habenicht et al. 2019).

5.3 Internal Structure

With regard to the empirical structure of the BilWiss 2–0 test, structural equation models indicated a good fit for the model with six correlated latent factors (for details, see Kunina-Habenicht et al. 2020). Moreover, we could show that educational knowledge can be invariantly measured across three subject groups, i.e., science, languages/humanities, and a combination of these subjects (Lohse-Bossenz et al. 2018).

5.4 Relation with Other Variables

Small significant correlations between the BilWiss 2.0 test score and the number of relevant university courses attended and grades in university studies support the convergent and prognostic validity of the test score interpretations. Moreover, teacher students who had to re-sit at least one exam in their educational foundation courses showed significantly lower test performance than students who passed on the first try (Kunina-Habenicht et al. 2020). Moreover, we found that university students with a range of individual risk factors such as lower cognitive abilities, lower SES, immigration background or unfavorable personality traits (e.g., high neuroticism scores) showed significantly lower test scores in later semesters than students without these risk factors (Wolf 2019).

6 Theoretical and Practical Implications

The BilWiss research program aimed to investigate the emergence and relevance of theoretical educational knowledge as an important facet of (prospective) teachers' professional competence. Over the period of almost ten years, we have succeeded to constructing and validating a psychometrically solid standardized test to assess generic educational knowledge in a generic, broad sense. A distinctive feature of the BilWiss test is that it not only includes those teaching-related knowledge areas that are typically targeted in existing concepts of pedagogical knowledge such as instruction, assessment, learning and development (Voss et al. 2015). The BilWiss test also covers topics that go beyond classroom teaching and touches other fields of teacher activity, such as school development, counseling, and an understanding of teaching as a professional occupation.

We found that teacher students and graduates of teacher education differ substantially in this form of knowledge which is at least partially due to a differing up-take of learning opportunities during teacher education, and that these differences in knowledge are associated with differences in professional success.

6.1 Implications for Future Research

After the comprehensive revision and validation of the first BilWiss test, the revised BilWiss 2.0 test will now be made available for other researchers⁴. It consists of a long version with 14–24 items per scale, providing a comprehensive assessment of most topics teacher education experts consider essential for prospective teachers, and a short version with 65 items that can be used as a condensed indicator for teachers' generic educational knowledge. We recommend the use of the long version in situations where one is interested to what degree an intended teacher education curriculum is realized or where one wants to assess selected dimensions of educational knowledge in more depth. Typically, this would be studies within the context of educational monitoring or evaluations of specific courses or treatments within the context of teacher education. We recommend the use of the short version in situations where one needs just a general individual score of educational knowledge as a measure of an important aspect of teachers' individual professional competence, figuring as a predictor, outcome or control variable.

The BilWiss 2.0 test can be applied for further research on the effectiveness and long-term impact of educational foundations courses. It is our conviction that this part of teacher education is better than its reputation and we expect other studies to confirm the high relevance of a sound theoretical base for successful mastery of the teaching job (Hollins 2011; Patrick et al. 2011).

We also hope that our test is a useful tool for studies investigating the development of educational knowledge through teacher education or specifically tailored interventions. Although there is ample research on effects of certain formats in teacher education, many of these studies suffer from a shortage of convincing objective measures (Cochran-Smith and Zeichner 2005). Moreover, the mechanisms underlying the development of teachers' professional knowledge are not well understood yet. For instance, it is not clear whether consecutive approaches in teacher education are more effective than concurrent approaches or vice versa (Harr et al. 2015). Another important open issue refers to the question of how theoretical knowledge can transfer into better practice. While it is unlikely that theoretical knowledge itself directly transfers into better teaching (or other professional) behaviors, it can be assumed that it plays an important role in the interpretation of teaching (or other professional) situations (Blömeke et al. 2015; König et al. 2014).

4 The BilWiss 2.0 test (in German language) will be made available at the Research Data Centre Education (FDZ Bildung) at the Leibniz Institute for Research and Information on Education (DIPF), the BilWiss data is available at the Research Data Centre (FDZ) at the Institute for Educational Quality Improvement (IQB).

Thus, the effects are not necessarily direct, but more complex and can only be recovered using moderation and/or mediation analysis (e.g., Dicke et al. 2015).

6.2 Implications for Practice

The general message from BilWiss research program is that the theoretical generic educational knowledge is one important condition for successful mastery of practice. Our findings show that gap between theory and practice may not be as wide as often purported: Not only do academic and practice experts agree which theoretical issues should be covered in initial teacher education. Moreover, our findings give first evidence that a good theoretical foundation can ease beginning teachers' entry into practice. Which forms of teacher education may be best suited to gain this type of knowledge remains an open question. With the BilWiss test, there is now a tool available with which new developments in teacher education may be empirically validated. The test is available for studies aiming at monitoring and evaluation of specific teacher programs. One essential issue for school policy and school administration is, for example, whether there is a difference between teachers who finished regularly certified teacher education programs, and lateral entry employees, who enter school after attending only reduced or even none teacher preparation programs and possess only limited knowledge about theoretical foundations. Although the BilWiss test was developed in German, we are convinced that its content can be at least partly adapted to other languages and countries.

References

- Abell, S. K. (2008). Twenty Years Later: Does pedagogical content knowledge remain a useful idea? *International Journal of Science Education*, 30(10), pp. 1405–1416. doi:10.1080/09500690802187041
- Alles, M., Apel, J., Seidel, T., & Stürmer, K. (2018). Candidate Teachers Experience Coherence in University Education and Teacher Induction: the Influence of Perceived Professional Preparation at University and Support during Teacher Induction. *Vocations and Learning*, 6(2). doi:10.1007/s12186–018-9211–5
- Bauer, J., Diercks, U., Rösler, L., Möller, J., & Prenzel, M. (2012). Lehramtsausbildung in Deutschland: Wie groß ist die strukturelle Vielfalt? [Teacher Education in Germany: How Big is the Structural Variety?]. *Unterrichtswissenschaft*, 40(2), pp. 101–120.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47, (pp. 133–180). doi:10.3102/0002831209345157

- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond dichotomies: Competence viewed as a continuum. *Journal of Psychology*, 223(1), 3–13. doi:10.1027/2151-2604/a000194
- Choy, D., Wong, A., Lim, K., & Chong, S. (2013). Beginning Teachers' Perceptions of their Pedagogical Knowledge and Skills in Teaching: A Three Year Study. *Australian Journal of Teacher Education*, 38(5). doi:10.14221/ajte.2013v38n5.6
- Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005). *Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education*. Washington, DC: American Educational Research Association.
- Council for Social Foundations of Education (CSFE). (1996). *Standards for Academic and Professional Instruction in Foundations of Education, Educational Studies, and Educational Policy Studies*. <http://www.unm.edu/~jka/csfe/standards96.pdf>. Access: 20 August 2019.
- Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in Teacher Preparation: How Well Do Different Pathways Prepare Teachers To Teach. *Journal of Teacher Education*, 53(4), pp. 286–302. doi:10.1177/0022487102053004002
- Dawson, D., Mazurek, K., & Deyoung, A. J. (1984). Courses in the social foundations of education: The students' view. *Journal of Education for Teaching: International Research and Pedagogy*, 10(3), pp. 242–248. doi:10.1080/0260747840100305
- Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education*, 34, pp. 12–25. doi:10.1016/j.tate.2013.03.001
- Dicke, T., Parker, P. D., Holzberger, D., Kunina-Habenicht, O., Kunter, M., & Leutner, D. (2015a). Beginning teachers' efficacy and emotional exhaustion: Latent changes, reciprocity, and the influence of professional knowledge. *Contemporary Educational Psychology*, 41, pp. 62–72. doi:10.1016/j.cedpsych.2014.11.003
- Dicke, T., Schmeck, A., Elling, J., & Leutner, D. (2015b). Reducing Reality Shock: The Effects of Classroom Management Skills Training on Beginning Teachers. *Teaching and Teacher Education*, 48, pp. 1–12. doi:10.1016/j.tate.2015.01.013
- EURYDICE. (2002). *The teaching profession in Europe: Profile, trends, and concerns. Report I: Initial training and transition to working life. General lower secondary education*.
- Grossman, P. L., & Richert, A. E. (1988). Unacknowledged knowledge growth: A re-examination of the effects of teacher education. *Teaching and Teacher Education*, 4(1), pp. 53–62. doi:http://dx.doi.org/10.1016/0742-051X(88)90024-8
- Guerrero, S. (Ed.). (2017). *Pedagogical Knowledge and the Changing Nature of the Teaching Profession*. Paris: OECD Publishing.
- Harr, N., Eichler, A., & Renkl, A. (2015). Integrated learning: ways of fostering the applicability of teachers' pedagogical and psychological knowledge. *Frontiers in Psychology*, 6(p. 738). doi:10.3389/fpsyg.2015.00738
- Hill, H. C., Rowan, B., & Loewenberg Ball, D. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal*, 42(2), pp. 371–406.
- Hohenstein, F., Zimmermann, F., Kleickmann, T., Köller, O., & Möller, J. (2014). Sind die bildungswissenschaftlichen Standards für die Lehramtsausbildung in den Curricula der Hochschulen angekommen? [Have the education standards for teacher training pro-

- grammes arrived in the university curriculum?] *Zeitschrift für Erziehungswissenschaft*, 17(3), pp. 497–507. doi:10.1007/s11618–014–0563–9
- Hollins, E. R. (2011). Teacher Preparation For Quality Teaching. *Journal of Teacher Education*, 62(4), pp. 395–407. doi:10.1177/0022487111409415
- Kennedy, M. M., Ahn, S., & Choi, J. (2008). The value added by teacher education. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre, & K. E. Demers (Eds.), *Handbook of research on teacher education* (3 ed., pp. 1249–1273). New York, NY: Routledge.
- König, J., & Pfanzl, B. (2016). Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. *European Journal of Teacher Education*, 39(4), pp. 419–436. doi:10.1080/02619768.2016.1214128
- König, J., Bloemeke, S., Klein, P., Suhl, U., Busse, A., & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education*, 38, pp. 76–88.
- Kunina-Habenicht, O., Maurer, C., Wolf, K., Holzberger, D., Schmidt, M., Dicke, T., Teuber, Z., Koc-Januchta, M., Lohse-Bossenz, H., Leutner, D., Seidel, T. & Kunter, M. (2020). Der BilWiss-2.0-Test: Ein revidierter Test zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften. *Diagnostica*. doi: 10.1026/0012-1924/a000238
- Kunina-Habenicht, O., Maurer, C., Schulze-Stocker, F., Wolf, K., Hein, N., Leutner, D., . . . Kunter, M. (2019). Zur curricularen Validität des BilWiss 2.0-Tests zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften [On the curricular validity of the BilWiss-2.0-Test for the assessment of (prospective) teachers' educational knowledge]. *Zeitschrift für Pädagogik*, 65(4), pp. 542–556.
- Kunina-Habenicht, O., Schulze-Stocker, F., Kunter, M., Baumert, J., Leutner, D., Förster, D., . . . Terhart, E. (2013). Die Bedeutung der Lerngelegenheiten im Lehramtsstudium und deren individuelle Nutzung für den Aufbau des bildungswissenschaftlichen Wissens [The Relevance of Learning Opportunities in Teacher Education Studies and their Individual Uptake for the Building of Educational Knowledge]. *Zeitschrift für Pädagogik*, 59(1), pp. 1–23.
- Kunina-Habenicht, O., Lohse-Bossenz, H., Kunter, M., Dicke, T., Förster, D., Gößling, J., . . . Terhart, E. (2012). Welche bildungswissenschaftlichen Inhalte sind wichtig in der Lehrerbildung? Ergebnisse einer Delphi-Studie [Which educational topics are important for teacher training? – Results of a Delphi study]. *Zeitschrift für Erziehungswissenschaft*, 15(4), pp. 649–682. doi:10.1007/s11618–012–0324–6
- Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology*, 78(2), pp. 75–95. doi:10.1037/0022–0663.78.2.75
- Linninger, C., Ewald, S., Stürmer, K., Seidel, T., & Kunter, M. (2016). *Educational knowledge: The groundwork for elaborate reflection?* Paper presented at the Conference of EARLI Special Interest Group 11 (Teaching and Teacher Education), Zürich, Switzerland.
- Linninger, C., Kunina-Habenicht, O., Emmenlauer, S., Dicke, T., Schulze-Stocker, F., Leutner, D., . . . Kunter, M. (2015). Assessing Teachers' Educational Knowledge: Construct Specification and Validation Using Mixed Methods. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 47(2), pp. 72–83. doi:10.1026/0049–8637/a000126

- Linninger, C., Kunina-Habenicht, O., Leutner, D., Seidel, T., Terhart, E., & Kunter, M. (2015). *Wer zeigt Engagement in der Schule? Individuelle Voraussetzungen proaktiven Verhaltens bei Lehrkräften [Who shows engagement in school? Individual prerequisites for teachers' proactive behaviors]*. Paper presented at the Fachgruppentagung Pädagogische Psychologie (PAEPS), Kassel.
- Linstone, H. A., & Turoff, M. (Eds.). (1975). *The Delphi method: Techniques and applications*. Cambridge, MA: Addison-Wesley.
- Lohse-Bossenz, H., Holzberger, D., Kunina-Habenicht, O., Seidel, T., & Kunter, M. (2018). Wie fach(un)abhängig ist bildungswissenschaftliches Wissen? – Messinvarianz und fachspezifische Unterschiede [How subject(un)-specific is educational knowledge? Measurement invariance and domain-specific differences]. *Zeitschrift für Erziehungswissenschaft*, 21 (5), pp. 991–1019. doi:10.1007/s11618–018-0817-z
- Lohse-Bossenz, H., Kunina-Habenicht, O., Dicke, T., Leutner, D., & Kunter, M. (2015). Teachers' knowledge about psychology: Development and validation of a test measuring theoretical foundations for teaching and its relation to instructional behavior. *Studies in Educational Evaluation*, 44, pp. 36–49. doi:10.1016/j.stueduc.2015.01.001
- Lohse-Bossenz, H., Kunina-Habenicht, O., & Kunter, M. (2013). The role of educational psychology in teacher education: Expert opinions on what teachers should know about learning, development, and assessment. *European Journal of Psychology of Education*, 28(4), pp. 1543–1565. doi:10.1007/s10212–013-0181–6
- Maurer, C., Ehlers, S., Wolf, K., Gartmeier, M., Hertel, S., & Kunter, M. (2018). *Teachers' professional vision in counseling situations: The development of a new assessment tool*. Paper presented at the Conference of EARLI Special Interest Group 11 (Teaching and Teacher Education) Kristiansand, Norway.
- Patrick, H., Anderman, L. H., Bruening, P. S., & Duffin, L. C. (2011). The Role of Educational Psychology in Teacher Education: Three Challenges for Educational Psychologists. *Educational Psychologist*, 46(2), pp. 71–83. doi:10.1080/00461520.2011.538648
- Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, 93(2), pp. 346–362.
- Rösler, L., Zimmermann, F., Bauer, J., Möller, J., & Köller, O. (2013). Interessieren sich Lehramtsstudierende für bildungswissenschaftliche Studieninhalte? Eine Längsschnittstudie vom ersten bis zum vierten Semester [Are teacher students interested in the content of educational foundation courses? A longitudinal study from the first to fourth semester]. *Zeitschrift für Pädagogik*, 59(1), pp. 24–41.
- Schleicher, A. (2016). *Teaching Excellence through Professional Learning and Policy Reform*. Paris: OECD Publishing.
- Schmidt, W. H., Cogan, L., & Houang, R. (2011). The Role of Opportunity to Learn in Teacher Preparation: An International Context. *Journal of Teacher Education*, 62(2), pp. 138–153. doi:10.1177/0022487110391987
- Schmidt, W. H., Houang, R., Cogan, L., Blömeke, S., Tatto, M., Hsieh, F., . . . Paine, L. (2008). Opportunity to learn in the preparation of mathematics teachers: its structure and how it varies across six countries. *ZDM – The International Journal on Mathematics Education*, 40(5), pp. 735–747. doi:10.1007/s11858–008-0115-y
- Schulze-Stocker, F. (2016). Die Normierung der Bildungswissenschaften: Wie reagieren lehrausbildende Universitäten in Nordrhein-Westfalen auf neue administrative Vorga-

- ben? [The standardization of educational foundation courses: How do teacher education universities in North Rhine Westfalia react to new administrative guidelines?] *Die Schule NRW*, 9, pp. 18–21.
- Schulze-Stocker, F., Holzberger, D., & Lohse-Bossenz, H. (2017). Das bildungswissenschaftliche Curriculum – Zentrale Ergebnisse des BilWiss-Programms. [The educational foundation curriculum – Key findings of the BilWiss program.] *Das Hochschulwesen*, 65(4+5), pp. 134–138.
- Schulze-Stocker, F., Holzberger, D., Kunina-Habenicht, O., Terhart, E., & Kunter, M. (2016). Spielen Studienschwerpunkte wirklich eine Rolle? Zum Zusammenhang von bildungswissenschaftlichen Studienschwerpunkten, selbsteingeschätzten Kenntnissen und gemessenem Wissen am Ende eines Lehramtsstudiums [Are educational foundation courses truly important? The relation between the course selection in educational foundations, self-assessed knowledge, and knowledge measured by a test at the end of university teacher education]. *Zeitschrift für Erziehungswissenschaft*, 19(3), pp. 599–623. doi:10.1007/s11618–016-0671–9
- Seidel, T., Stürmer, K., Prenzel, M., Jahn, G., & Schäfer, S. (2017). Investigating pre-service teachers' professional vision within university-based teacher education. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education – research, models and instruments* (pp. 93–110). New York: Springer.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), pp. 4–21.
- Sonmark, K., Révai, N., Gottschalk, F., Deligiannidi, K., & Burns, T. (2017). *Understanding Teachers' Pedagogical Knowledge: Report on an International Pilot Study. OECD Education Working Papers* (Vol. 159). Paris: OECD Publishing.
- Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). (2014). Standards für die Lehrerbildung: Bildungswissenschaften [Standards for teacher education: Educational foundation studies]. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf. Access: 29.07.2019.
- Stürmer, K., & Seidel, T. (2015). Assessing Professional Vision in Teacher Candidates. *Zeitschrift für Psychologie*, 223(1), pp. 54–63. doi:10.1027/2151–2604/a000200
- Terhart, E., Lohmann, V., & Seidel, V. (2010). *Die bildungswissenschaftlichen Studien in der universitären Lehrerbildung. Eine Analyse aktueller Studienordnungen und Modelhandbücher an Universitäten in Nordrhein-Westfalen [The educational foundation courses in university teacher education. An analysis of current study regulations at universities in North Rhine Westfalia]*. Unpublished report. Institute of Educational Science. Westfälische Wilhelms-Universität Münster.
- Tozer, S. E., & McAninch, S. (1986). Social Foundations of Education in Historical Perspective. *Educational Foundations*, 1(1), pp. 3–32.
- Veenman, S. (1984). Perceived problems of beginning teachers. *Review of Educational Research*, 54(2), pp. 143–178. doi:10.3102/00346543054002143
- Voss, T., Kunina-Habenicht, O., Hoehne, V., & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften: Empirische Zugänge und Befunde [Keyword: Teachers' Educational Knowledge: Empirical Approaches and Findings]. *Zeitschrift für Erziehungswissenschaft*, 18(2), pp. 187–223. doi:10.1007/s11618–015-0626–6

- Voss, T., Kunter, M., Seiz, J., Hoehne, V., & Baumert, J. (2014). Die Bedeutung des pädagogisch-psychologischen Wissens von angehenden Lehrkräften für die Unterrichtsqualität [*Zeitschrift für Pädagogik*, 60(2), pp. 184–201.
- Wilson, S., & Floden, R. E. (2003). *Creating Effective Teachers: Concise Answers for Hard Questions. An Addendum to the Report "Teacher Preparation Research: Current Knowledge, Gaps, and Recommendations"*. <https://files.eric.ed.gov/fulltext/ED476366.pdf>. Access: 29 July 2019.
- Wilson, S., Floden, R. E., & Ferrini-Mundy, J. (2002). Teacher preparation research: An insider's view from the outside. *Journal of Teacher Education*, 53(3), 190–204. doi:10.1177/0022487102053003002
- Wilson, S., Floden, R. E., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations, A research report prepared for the U.S. Department of Education*. <https://www.education.uw.edu/ctp/sites/default/files/ctpmail/PDFs/TeacherPrep-WFFM-02-2001.pdf>. Access: 29 July 2019.
- Wolf, K. (2019). *Individuelle Unterschiede auf dem Weg zur Lehrkraft: Die Bedeutung persönlicher Eingangsvoraussetzungen für den Studien- und Berufserfolg von (angehenden) Lehrkräften [Individual differences in the route towards becoming a teacher: The relevance of personal prerequisites for the study and job success of (future) teachers]*. (PhD thesis, Goethe Universität Frankfurt am Main, Germany).
- Wong, A. F., Chong, S., Choy, D., Wong, I. Y., & Goh, K. C. (2008). A Comparison of Perceptions of Knowledge and Skills Held by Primary and Secondary Teachers: From the Entry to Exit of Their Preservice Programme. *Australian Journal of Teacher Education*, 33(3). doi:10.14221/ajte.2008v33n3.6
- Zeichner, K. M. (2006). Reflections of a University-Based Teacher Educator on the Future of College- and University-Based Teacher Education. *Journal of Teacher Education*, 57(3), pp. 326–340. doi:10.1177/0022487105285893
- Zeichner, K. M. (2005). A research agenda for teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The AERA Panel on Research and Teacher Education* (pp. 737–760). Mahwah, N. J.: Lawrence Erlbaum.



2.2

Analyses and Validation of Central Assessment Instruments of the Research Program TEDS-M

Kaiser, G., and König, J.

Abstract

The TEDS-Validate project has been carried out within the research program derived from Teacher Education and Development Study: Learning to Teach Mathematics (TEDS-M). In this chapter, we describe the aim of the study, which is related to the question of whether research findings brought forward by measurement instruments to test professional competence of mathematics teachers have predictive validity for the quality of their classroom instruction and the learning gains of their students. Based on this, we question whether situation-specific skills (measured via video-based assessments) contribute to explain instructional quality and learning gains of students – in addition to the effects of professional knowledge of teachers. To answer the research questions, data was collected in Thuringia, Saxony and Hesse from 2016 to 2019 with a survey of 113 in-service teachers. They were tested using web-based instruments to capture their mathematics, mathematics pedagogical and general pedagogical knowledge as well as their noticing competencies. TEDS-Validate points out the broad applicability of the instruments for the comprehensive measurement of mathematics teachers competencies. To evaluate the effects of prospective teachers' practical activities during their school practicum within the master study of initial teacher education, these instruments will be applied in a follow-up transfer project.

Keywords

Mathematics teachers' competencies, teachers' knowledge, noticing, instructional quality, students learning gains, international comparative studies, transfer activities

1 Introduction

TEDS-Validate has been carried out within the research program, which has been derived from the international comparative study on teacher education, the *Teacher Education and Development Study: Learning to Teach Mathematics (TEDS-M)* (2008–2010) (Blömeke et al. 2010a, b). The design, conceptualisation and instruments of TEDS-Validate refer to the TEDS-M research program and has evolved in the last ten years.

1.1 Aims of the TEDS-Validate Project

The TEDS-Validate project aims to answer the question of whether research findings brought forward by TEDS-M measurement instruments to test professional competence of mathematics teachers have predictive validity for the quality of their classroom instruction and the learning gains of their students. The used instruments have been developed within the study TEDS-M and measure the professional competencies of prospective mathematics teachers. It is assumed that if predictive validity of the instruments of teachers' competencies on quality-oriented teaching and students' learning gains can be confirmed, then these assessment instrument can be considered objective, reliable, and valid and enable examination of teacher effectiveness in future research. The measurement instruments developed within TEDS-M, which consisted of teachers' knowledge instruments and capture the central cognitive elements of mathematics teachers' professional competence (Shulman 1987, Baumert and Kunter 2006), were already validated: mathematical content knowledge (MCK), mathematical pedagogical content knowledge (MPCK), and general pedagogical knowledge (GPK). However, it is still an open question whether these instruments fulfil predictive validity concerning the mastering of professional tasks relevant during teaching as the core task of teachers.

Another objective of the TEDS-Validate project was to analyze which kinds of teacher knowledge – declarative (“knowing that...”) or procedural (“knowing how...”) – are acquired during initial teacher education and teachers' profession-

al practice. The original TEDS-M measurement instruments (e.g., Buchholtz et al. 2016; König and Blömeke 2010) were mainly focused on declarative cognitive knowledge, which is acquired during initial teacher education and are an essential part of teacher's knowledge for teaching. Procedural knowledge relies strongly on practical experience, being more related to teaching situations and performance in class. This part of the overall teachers' competencies was evaluated by specific further development of the original TEDS-M-instruments, namely video-based assessment from the project TEDS-Follow-up (TEDS-FU) (e.g., Blömeke et al. 2014; König et al. 2014). These innovative assessment instruments were evaluated by expert reviews and have been proven to be reliable and valid as well as being suitable for capturing situation-specific skills. The additional question within TEDS-Validate was whether these instruments also have predictive validity, which would allow to gain deeper insight for the design of initial teacher education.

The international discourse on professional competence developed during initial teacher education or professional development has strong gaps. Although a chain of effects such as teacher preparation – teacher competence – instructional quality – students' learning gains is generally assumed (e.g. Baumert et al. 2010, Hill et al. 2005), there is only little empirical evidence whether competencies teacher acquire during initial teacher education have an influence on the quality of their instruction and the learning gains of their students. Especially the relationship of teacher sub-facets of their competencies such as CK, PCK, and GPK on instructional quality and student learning has not yet been modelled simultaneously, neither have situation-specific cognitive skills been accounted for.

1.2 Research Questions

In the TEDS-Validate project, two major research questions were guiding the empirical study:

1. How far can we provide empirical evidence that the measurement instruments developed within the TEDS-M and TEDS-FU projects have predictive validity for teaching mathematics at high quality?

We expect that MCK, MPCK, and GPK as well as video-based measures significantly impact instructional quality and correlate with the learning gain of students.

2. Do situation-specific skills (measured via video-based assessments) contribute to explain instructional quality and learning gains of students – in addition to

the effects of professional knowledge of teachers (as measured via paper-pencil tests), i.e., do they have added value?

We expect that video-based skills of perception, interpretation, and decision-making correlate higher with instructional quality and student learning gains than knowledge-based tests (MCK, MPCK, GPK). Moreover, we expect that the relationship between video-based measures and students' learning gains will be mediated by instructional quality.

2 Literature Review

2.1 Teacher Competence and Its Measurement

The concept of competence in educational research has been developed within different research traditions. The broad definition developed by Weinert (2001), which incorporates (i) cognitive abilities, (ii) the motivation, volition, social willingness, and ability to solve problems, and (iii) the motivation, volition, and social readiness to implement solutions, has shaped the discussion of competence in large-scale studies on teacher education (Kunter et al. 2011; Blömeke et al. 2010a, b). Klieme and Leutner (2006), who define competencies as situation-specific, cognitive performance dispositions, which are functionally responsive to situations and demands in specific domains, have further developed this work. Within the discussion of professionalization of teachers, generic models of professional competence are currently proposed encompassing cognitive and affective-motivational aspects (Baumert and Kunter 2006; Blömeke 2017). The differentiation of teachers' dispositions into various knowledge facets by the seminal work of Shulman (1987) has shaped the scientific discourse until today (Guerrero 2017). In the last decade, this approach has been widened going beyond the description of teacher competencies as personal traits (i.e., individual dispositions relatively stable across different contexts) including situational facets (Kaiser et al. 2017).

Similar developments can be identified in subject-related discussions on teacher education. For example, in mathematics education, Krainer and Llinares (2010) described various trends in the literature about prospective teachers, practicing teachers, and teacher educators, amongst others reflecting a shift from the individual perspective on teachers toward emphasizing the social dimension in teacher education based on sociological and sociocultural theories. In connection with this shift towards the social dimension different paradigms on teachers' professional competencies have been identified by Rowland and Ruthven (2011), which can

be characterized either as cognitive or as situated approaches to the professional competencies of teachers (for an extensive overview on these paradigmatic distinctions, see Kaiser et al. 2017).

This development towards the social dimension on the professional activities of teachers exhibits the transfer from a cognitive perspective focusing on the knowledge facets of teachers within teacher professionalism to situated approaches. The cognitive perspective has been dominant in recent decades and is characterized by its focus onto a limited number of components related to personal traits. Prominent examples of large-scale studies mainly come from mathematics education, such as the *Teacher Education and Development Study in Mathematics* (TEDS-M) or the *Professional Competence of Teachers, Cognitively Activating Instruction and the Development of Students' Mathematical Literacy* (COACTIV). Follow-up studies of these and other studies consider the multidimensionality of teacher competencies and include context-specific and situated aspects of teaching and learning. Especially the concept of teacher noticing plays an important role. In the newly developed framework of teacher competencies Blömeke et al. (2015) show that older conceptual dichotomies ignore either the stable dispositional or the more variable situational competence facets. According to this model, competencies can be described along a continuum from personal dispositions, namely teacher professional knowledge and beliefs, which are complemented by situation-specific cognitive skills such as perception, interpretation, and decision-making, which finally lead to teacher performance in the classroom.

The research model used in the project TEDS-Validate (Section 3.2.1), refers to this theoretical approach describing professional teacher competencies as blend of cognitive and affective-motivational dispositions as well as situational-specific skills (Blömeke et al. 2015, Kaiser et al. 2017). The different paradigms on the conceptualization of teacher competencies have consequences regarding competence measurement. On the one hand, classical paper-and-pencil tests capturing the different facets of teacher knowledge have been developed, primarily in mathematics education but also within other domains (Großschedl et al. 2015 for biology education, König et al. 2016 for education of English as a foreign language, Krauss et al. 2017 for mathematics, German, English, Latin, physics, music, religious education; see also contributions in this volume). On the other hand a number of research groups referring to teacher noticing or professional vision has developed video-based assessment instruments (e.g., Kersting et al. 2012; Seidel and Stürmer 2014, Kaiser et al. 2015; for an overview of the construct of noticing, see Sherin et al. 2011, more recent Schack et al. 2017, a systematic literature review on the measurement of noticing is provided by Stahnke et al. 2016).

2.2 Research Desiderata Addressed by TEDS-Validate

Studies that analyze teachers' certificates and qualifications in terms of their impact on student learning (e.g., Darling-Hammond et al. 2001; Cochran-Smith and Zeichner 2005; Palardy and Rumberger 2008) have certain limitations. They have been criticized for the fact that the teacher quality indicators used do not sufficiently explain variations in the quality of their teaching or the learning progress of their students. Therefore, researchers have started to design studies that directly assess teacher knowledge, and the indicators for professional competence of teachers are examined, for instance, regarding its relationship to instructional quality and student learning (Hill et al. 2005; Baumert et al. 2010; Kersting et al. 2012).

Measurement instruments for cognitive facets of teacher professional competence have been developed over the past decade with particular focus on mathematics. In Germany, the research study COACTIV (Kunter et al. 2011) is well known, however, its findings are based on data collected 15 years ago. PISA 2003 was extended by a second time point one year later with another student assessment in mathematics, allowing the analysis of student progress in year 9 and 10. Their teachers were assessed using paper-pencil tests measuring CK and PCK and measures were applied to capture the quality of mathematics instruction (Baumert et al. 2010). The COACTIV study proliferated evidence that teachers' CK and PCK effect instructional quality (such as cognitive activation) and influence student learning in mathematics. More recently, data on teachers' noticing skills were related to their knowledge base (Bruckmaier et al. 2016). These findings relate to the international state of art on CK and PCK as brought forward by Hill et al. (2005) in the US context. In the COACTIV-R study in Germany, empirical evidence was provided for the effect of pedagogical-psychological knowledge of pre-service teachers during induction on the instructional quality of classroom management as perceived by students (Voss et al. 2014). In a study conducted in Austria, König and Pflanzl (2016) provided evidence for the effect of TEDS-M test measuring GPK on student perceptions among about 250 in-service teachers. If teachers performed well on the GPK test, students reported higher instructional quality regarding effective classroom management, teacher clarity, and positive teacher-student relationships, even when controlled for teacher personality (Big-Five), teaching experience, and teacher certification grades.

Although these findings are promising towards teacher competence and its relevance for teaching and student learning, there is still a lack of research:

- First, nearly all studies basically assume a chain of effectiveness related to 'teacher education – teacher competence – instructional quality – student learn-

ing'. Although numerous studies have shown pre-service teachers may acquire competencies (e.g., teacher knowledge) during initial teacher education (for a recent overview, see Kaiser and König 2019), hardly any evidence exists as to what extent these teacher competencies acquired during initial teacher education at higher education institutions have an impact in the long run. That means it is still unclear whether teacher competencies that represent an outcome of higher education actually are significant predictors for instructional quality teachers provide during in-service teaching and the progress of their students.

- Second, none of the studies has analyzed the triad of CK, PCK, and GPK in an overall model. So hardly any insight can be given into the multidimensionality of teacher competence and its specific impact on instructional quality and student learning when modelling the triad simultaneously.
- Third, due to recent developments in competence modelling, not only teacher knowledge as a cognitive disposition, but also situation-specific teacher skills have to be considered (Blömeke et al. 2015, Kaiser et al. 2015, 2017). Only very few studies have started to reflect on such a distinction in the complex field of teacher competence measurement (e.g., Blömeke et al. 2016, Kersting et al. 2012, König et al. 2014). These studies again focus on either subject-specific or generic facets of teacher competence, so they have the limitation that they cannot include all facets in an overall statistical model.

TEDS-Validate specifically focuses on these research desiderata. The whole set of measurement instruments developed in the TEDS-M research program, including TEDS-M, TEDS-FU, TEDS-Instruct, have been applied. Therefore, its findings should have relevant implications for teacher education, as teacher education programs usually consist of the typical components related to the subject, subject-specific pedagogy, and general pedagogy (Flores 2016). The differentiation into teacher knowledge and situation-specific skills as cognitive elements of the professional competence of teachers (Figure 1) reflects the ongoing discussion on theory-practice in teacher education, which is of great relevance in the current German Quality Initiative of Teacher Education ([*Qualitätsinitiative Lehrerbildung*], BMBF 2014).

3 Design of the Study

3.1 Overview of the TEDS-M Research Program

The *Teacher Education and Development Study in Mathematics* (TEDS-M) carried out under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) was a comparative study of teacher education and the first IEA study on tertiary education, as well as the first international large-scale assessment of future teachers that based on representative samples (Tatto and Senk 2012). The TEDS-M target population consisted of mathematics teachers for elementary and secondary schools in their final year of teacher education. Data were collected in 2008. A central component of TEDS-M was the measurement of the professional knowledge of prospective teachers. The common international questionnaire only measured prospective teachers' mathematics content knowledge (MCK) and mathematics pedagogical content knowledge (MPCK). Three participating countries – the United States, Germany, and Taiwan – therefore carried a national version measuring prospective teachers' GPK. Besides pre-service teacher knowledge, also, their beliefs were investigated and a broad range of institutional characteristics of teacher education programs such as learning opportunities as well as socio-demographic variables were part of the prospective teacher surveys (for details, see Blömeke et al. 2010a, b, 2014).

TEDS-M had a particular impact on starting empirical research on teacher education in Germany. As a consequence, several studies were conducted that systematically built on the TEDS-M research: TEDS-Follow Up (TEDS-FU), TEDS-Learning to Teach (TEDS-LT), and TEDS-Instruct (TEDS-U) building the comprehensive TEDS-M research program (Table 1).

While the TEDS-M study intended to evaluate the efficiency of teacher education at an international level by assessing prospective mathematics teachers' content-related knowledge at the end of teacher education, including the study courses and opportunities to learn, the other TEDS-M-studies had different objectives, respectively. They all were conducted in Germany or in specific federal states of Germany and they focused on different target groups.

Table 1 Studies of the TEDS research program

	TEDS-M	TEDS-LT	TEDS-FU	TEDS-U	TEDS-V
		(Learning to Teach)	(Follow-Up)	(Instruct)	(Validate)
	2006–2010	2008–2012	2010–2013	2015–2018	2016–2019
Geographical focus	international	Bavaria, Baden-Wuerttemberg, Berlin, Hamburg, Hesse, North Rhine-Westphalia	Germany	Hamburg	Hesse, Saxony, Thuringia
Target group	Pre-service teachers in their final year of teacher education	Future teachers at universities	Early career teachers	In-service teachers	
Subject	Mathematics	German, English as a Foreign Language, Mathematics	Mathematics		
Teaching Type	Primary and Secondary	Secondary	Primary and Secondary	Secondary	

- TEDS-LT aimed at transferring the conceptualizations and assessment approaches of TEDS-M to other domains such as German and English as a Foreign Language. TEDS-LT was continued in two more projects on language teacher education: PKE (König et al. 2016) and PlanvoLL-D (see the chapter by König et al. in this volume).
- TEDS-FU as a follow-up study of TEDS-M was more directly linked to TEDS-M, since it re-examined teacher professional competence of those participants who had actually entered the teaching profession four years later. TEDS-FU enriched the knowledge oriented TEDS-M-study by context-specific and situated aspects of teaching and learning including the concept of teacher noticing. TEDS-FU examined the assumption of the multidimensionality of teacher competencies of referring not only to subject-based cognitive aspects but also to pedagogical reflections on the teaching-and-learning situation as a whole. Overall, the context in which teaching and learning are enacted was therefore been brought to the foreground by TEDS-FU.

- TEDS-Instruct (TEDS-Unterricht, TEDS-U) was conducted as a pilot study exploring conceptualizations, instruments, and analyses approaches to be validated in TEDS-Validate. Mathematics teachers recruited for TEDS-Instruct (convenience sample due to data collection constraints) either taught at the academic track (*Gymnasium*) or at the non-academic track (*Stadtteilschule*) in the federal state of Hamburg. Classroom observations with a newly developed instrument were carried out with a subsample of the studied teachers to evaluate the instructional quality of their teaching. For each school class of the participating teachers, student assessment data were made available by the Institute for Educational Monitoring and Quality Development in Hamburg.

3.2 Description of TEDS-Validate: Research Model, Sample, and Instruments

TEDS-Validate builds on previous work done in the TEDS-M research program, in particular, it uses measurement instruments developed in the context of these previous research studies. These were analysed and the validity of these instruments were examined.

To investigate the two major research questions (Section 1.2), a specific research model was developed (Figure 1). It is grounded in relevant paradigms prominent in current empirical educational research such as effective teaching (e.g., Hattie 2009; Helmke 2012) and teacher expertise (Berliner 2004; Stigler and Miller 2018). Referring to current research on instructional quality such as the concept of generic dimensions being relevant for student learning, we included cognitive activation, constructive support, and effective classroom management and extended this framework by subject-specific dimensions on mathematics instructional quality. One basic assumption is that these dimensions of instructional quality significantly contribute to students' learning gains, mediated by learning processes triggered by the teachers. To describe teachers and their role in the teaching-learning-interaction, a particular focus of TEDS-Validate was on a complex model of cognitive elements of professional competence of teachers. Following current ideas in competence modeling by Blömeke et al. (2015), we differentiated into teacher knowledge as being distal cognitive elements and video-based measures of situation-specific skills as being more proximal cognitive elements for teaching. These core elements – teachers, teaching, and students – were framed by background and demographics of both teachers and students as well as context variables such as class composition characteristics or school type.

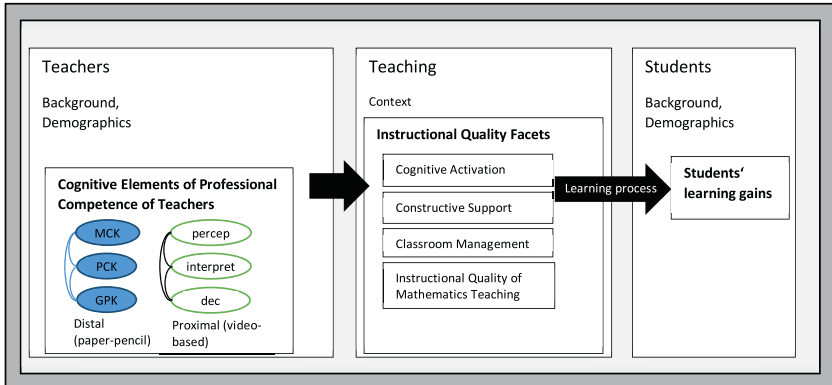


Figure 1 TEDS-Validate research model

TEDS-Validate has been conducted in Thuringia, Saxony and Hesse from 2016 to 2019, 113 in-service teachers participated in the study, 70 (62 %) in Thuringia, 13 (11 %) in Saxony and 30 (27 %) in Hesse. 64 participants (57 %) were female, their average grade in the first state examination was 2.1. (out of a scale ranging from 1 to 4 with 1 indicating highest level and 4 indicating just having passed) with one being the best grade and 1.9 in the second state examination. Their average age was 43 years with a range from 26–64, accordingly their professional experience averaged to 16 years with a range from 0.5–40 years. The evaluation of cognitive and situation-specific tests took 2.5 hours testing time and was carried out online at home by the teachers (using the survey software package and provider ‘Unipark’), usually with a break between the two test parts. The teachers in Thuringia received financial incentives in addition to courses for professional development, in Hesse and Saxony only professional development courses as incentives were permitted.

Moreover, instructional quality was evaluated with a sub-group of teachers ($n = 39$), 20 (51 %) in Thuringia, 5 (13 %) in Saxony and 14 (36 %) in Hesse. 23 participants (59 %) were female, their average grade in the first and second state examination was nearly identical to the whole sample (2.0 and 1.9). Their average age and range was very similar to the whole sample (44 years and 26–64), the same held for their professional experience (average 17 years and a range from 1–40 years). The teachers were observed within their teaching two times by two raters using the newly developed rating instrument (Jentsch et al. under review; Schlesinger et al. 2018). The University of Jena provided the data on students’ learning gains through the project ‘kompetenztest.de’. Students’ learning gains

have been measured via central assessments that regularly take place during grade 6 and 8 in Thuringia and Saxony and which are aligned with the national educational standards for mathematics (<https://www.kompetenztest.de/>). These data could not be collected in Hesse as no second testing takes place in this federal state.

As TEDS-Validate aimed at proliferating innovative approaches of measuring teacher competence the following instruments developed within the TEDS-M-research program have been used in the study:

- Online measurement instruments for MCK, MPCK, and GPK as the relevant categories of teacher professional knowledge originally stemming from TEDS-M and transferred from paper-pencil-format into online-format (Blömeke et al. 2010a, b, 2014);
- Online instrument on Perception, Interpretation and Decision-Making (PID) in the mathematics classroom using three video-vignettes related to mathematics pedagogy, general pedagogy (Kaiser et al. 2015), and an additional video-based instrument on effective classroom management expertise of teachers (CME, König 2015; König and Kramer 2016);
- Online test on the ability to quickly identify student errors (Pankow et al. 2016);
- In-vivo evaluation of instructional quality using a novel observational instrument (Schlesinger et al. 2018); in addition 15 of the teachers of TEDS-Instruct were video graphed within their original classes to compare the effect of in-vivo-coding and video-based coding (Benecke 2018).

4 Results¹

In the framework of the TEDS-M research program TEDS-Validate focused on the validation of the conceptualizations and instruments developed within TEDS-M, especially on the predictive validity for teaching mathematics at high quality and the impact of situation-specific sub-competencies of teachers on instructional quality and students' learning gains. Overall, the breadth of the TEDS instruments inventory allowed a complete operationalization of novel understanding of teacher competence as being a continuum comprising cognitive dispositions, situation-specific skills, and performance (Blömeke et al. 2015; Kaiser et al. 2017).

1 In the following core results of TEDS-Validate are presented, although the publication process is still in strong progress.

4.1 Construct Validation of Pedagogical Knowledge

Pedagogical knowledge of teachers is a relevant category of the professional teacher's knowledge base (Shulman 1987; König 2019). The test instrument developed in TEDS-M turned out to be valid for comparative assessment in the US, Germany, and Taiwan (König et al. 2011). The test focuses on pedagogical tasks teachers have to master such as classroom management, dealing with student heterogeneity, or assessment, whereby the knowledge focused on in the test is conceptualised as link to a specific subject. The TEDS-M test has proven to be reliable for measuring the knowledge of pre-service and in-service teachers, based on several studies, and broad evidence of its validity across professional education phases and contexts has been provided (for an overview, see König 2014).

In TEDS-Validate, the examination of the test's construct representation (Embretson 1983) was studied more carefully. The analyses focused on the question whether the test also requires and represents construct-relevant cognitive processes of professionally experienced mathematics teachers. Based on the data of TEDS-Validate, it was examined whether task difficulty could be attributed to cognitive processes as suggested in a proficiency model proposed by König (2009; in addition Klemenz and König 2019). The findings point out that the measurement is reliable among the TEDS-Validate target group of in-service teachers (EAP/PV-reliability .74). Moreover, the complexity level of task solutions explained a significant proportion of the variance in item difficulty (*adj. R*² = .48). Overall, first evidence could be provided for pedagogical knowledge construct representation. Furthermore, teachers' situation-specific skills could be predicted by the complexity level. This result can be interpreted as further proof of validation of the TEDS-M instrument measuring pedagogical knowledge (for details, see Nehls et al. under review).

4.2 Profiles of Teachers

Profiles of teachers were examined in two analyses based on data of TEDS-Validate, which aimed to study the relation between various competence facets of teachers.

In the first study pedagogical knowledge was focused on in a specific analysis to investigate profiles of teachers. Starting from the assumption that qualitatively different profiles of teacher pedagogical knowledge exist, a Mixed-Rasch model was applied to identify such profile groups of teachers. For these analyses, a larger data base including several TEDS-M-studies was created: TEDS-Validate data

was merged with data from TEDS-FU and CME, which is another study on in-service teachers in Germany that applied the GPK test from TEDS-M (König 2015). Therefore, a total sample of 462 in-service teachers, mathematics and non-mathematics teachers was used, who had completed the GPK test. 35 teachers (8 %) had undergone training for teaching mathematics at primary school level, 343 (74 %) for mathematics at secondary school level, while 81 (18 %) had been trained for other subjects at secondary school level. The participants had an average teaching experience of about 11 years ranging from 0.5 to 41.0 years. Two kind of analyses were carried out, mixed Rasch models with increasing number of classes and a study of the resulting GPK profiles at item-level.

The profile analysis released two different groups: The two profiles differed in their overall test performance (quantitative differences in test scores), and they also showed differences related to the quality of responding to single items, which resulted in a variant ranking of difficulty in some items. These items were mainly related to the pedagogical task of dealing adaptively with student heterogeneity in the classroom. Teachers with qualification in mathematics outperformed non-mathematics teachers with regard to these items specifically.

Further validation analysis was carried out towards teacher beliefs (epistemological beliefs and beliefs of teaching and learning) and instructional quality those teacher groups provided to their students, confirming the higher competence level of the teacher profile group who was more successful in the GPK test (for details, see Nehls et al. 2020).

The second study focuses the relation between instructional quality implemented by mathematics teachers' and those competences. Although there exists quite a significant number of studies already, most studies have up to now been restricted to a limited set of constructs and variable-oriented approaches that assume sample homogeneity. As in TEDS-Validate teacher competence is conceptualised as a comprehensive multi-dimensional construct including subject-specific and generic facets regarding knowledge, skills and beliefs, these analyses followed a different approach. In contrast, the used person-oriented approach examined whether subgroups of teachers exist when teachers are assessed on a large set of competence facets. The analyses assessed mathematics teachers' subject-specific and generic knowledge, skills and beliefs applying latent profile analysis to this broad range of teacher measures. These profiles were then related to instructional quality implemented in terms of mathematics teaching. Based on data from TEDS-Instruct and TEDS-Validate four groups of mathematics teachers could be distinguished, who differed quantitatively and qualitatively. The first group of teachers with pronounced levels of knowledge and skills succeeded with respect to student support, cognitive activation and mathematics-related quality while they implemented a

medium level of classroom management. The second group of teachers with high levels of cognitive skills in professional noticing under a mathematics educational perspective (M_PID) and classroom management succeeded with respect to cognitive activation and mathematics-related quality despite only medium levels of knowledge. Classroom management was a strength of a third group of teachers characterized by medium levels of all competence facets. These teachers struggled with mathematics-related quality, cognitive activation and student support though. The fourth small group of teachers with rather low levels of knowledge and skills struggled with all facets on instructional quality (for details, see Blömeke et al. accepted).

4.3 Results Concerning Instructional Quality: Validation of the Instrument and Relations to Teachers' Competencies

A new instrument for the evaluation of instructional quality was developed within the framework of the TEDS-M research program based on live ratings of the instructional quality classroom teaching. Based on data of TEDS-Validate and TEDS-Instruct the three basic dimensions of instructional quality (classroom management, student support and cognitive activation) and a content-specific dimension referring to mathematics-educational aspects (e.g. use of representations, explanations and examples) were measured (Jentsch et al. under review). Moderate correlations between teachers' professional noticing and domain-specific dimensions of instructional quality could be identified, but only weak correlations between PCK and subject-related quality of instruction were observed (Schlesinger et al. 2018). Overall analyses on the impact of teacher competencies on instructional quality and students' learning gains are currently in progress (König et al. under review).

To validate the instrument on teaching quality used for class observation, a supplementary project to TEDS-Validate, the TEDS-Video study, was carried out in which 15 teachers of the TEDS-Instruct study from Hamburg were subjected to renewed class observation and these lessons were simultaneously videographed. The sample consisted of 8 (53 %) females, grades in first and second teacher examination were 1.8 and 2.0, and the average age was 36 years ranging from 28 to 71. The professional experience averaged to 6 years ranging from 0.5 to 30 years. A comparison of the methods between live rating directly in the classroom and video rating from the videographed lesson was made possible by a subsequent rating of these videos after live rating had taken place by different raters. The results available so far indicate that mode effects only occur in some dimensions of

teaching quality due to the two types of rating. Video rating yielded slightly higher rank orders in classroom management than live rating ($z = -2.67, p < .01$, two-sided Wilcoxon test). We also found that student support was rated more reliable through video analyses ($\rho = .20$ vs. $\rho = .65$), whereas cognitive activation received better psychometric results during live rating ($\rho = .73$ vs. $\rho = .51$) (Cronbach et al. 1972). There was no meaningful difference between modes for the content-specific dimensions. Since the videographies also offer the possibility to analyse the lessons qualitatively, analyses on interaction patterns in the lessons are planned, especially on how teachers deal with student errors in their lessons (Benecke 2018).

4.4 Instructional Quality Measured via Task Quality

In the COACTIV study, instructional quality was measured via the quality of the mathematical tasks used by the teachers. Additional analyses refer to this approach and further develop the classification system of task quality based on a rational task analysis (Jordan et al. 2006). This classification system is currently applied to analyse the quality of the tasks used by 31 teachers from TEDS-Validate, who took part in the classroom observation study on instructional quality. Altogether 2600 mathematical tasks, i.e. between 25 and 197 tasks per teacher (two lessons of 90 minutes) form the data base and are analysed according to content-related and procedural competencies, cognitive and linguistic complexity and special task characteristics. The analyses aim to examine the central assumption that there is a strong correlation between instructional quality measured via classroom observations and task quality (for the first results, see Ross and Kaiser 2018).

4.5 Comparative Study between East and West

In a complementary study – the study TEDS-East-West – the question was examined how far theoretical frameworks, conceptualisations and instruments on teacher competencies developed in a Western context can be transferred to an East Asian context using China and Germany as paradigmatic examples of East and West. Analyses pointed out major similarities of the two contexts within in the conceptualization of teacher competence as a multidimensional construct comprising knowledge, situation-specific sub-competencies, and beliefs concerning the teaching and learning of mathematics. Distinct differences could be identified with the Chinese frameworks putting more teaching-related competencies in the foreground.

Extensive adaptation and validation studies were carried out using data from TEDS-Instruct (118 teachers) and newly sampled data in China (203 teachers). A comparative analysis of teacher competence frameworks developed in Eastern (Chinese) and Western (German) contexts pointed out major similarities of the two contexts, for example the conceptualization of teacher competence as a multi-dimensional construct comprising knowledge, teaching-related skills, and beliefs. Distinct differences could be identified as well, with the Chinese frameworks emphasizing more teaching-related competencies than the Western (German) frameworks. Based on a qualitative approach on examining validity of the framework and the instruments used, namely elemental validity and a quantitative approach, namely construct validity to validate the framework, the results of both approaches suggest satisfactory validity for the adaptation. Overall, the results pointed out that the examined teacher competence framework and its instruments can be used for comparative analyses in both countries (Yang et al. 2018).

Further comparative analyses on teachers professional noticing, i.e. perception, interpretation, and decision-making competencies, which was evaluated video-based, pointed out that German teachers showed significantly better achievements than Chinese teachers on noticing aspects related to general pedagogy. Chinese teachers performed more strongly than their German counterparts on noticing aspects connected to mathematics instruction. Further analysis found that German mathematics teachers showed particular strengths in the sub-facet perception of noticing, whereas Chinese teachers tended to be strong in the sub-facet ‘analysing and decision making’ of central classroom incidents. These results pointed out that societal and cultural factors, such as different philosophical paradigms, traditions of teacher education, and teaching and mathematics curriculum traditions are influencing strongly teachers’ professional noticing (Yang et al. 2019).

4.6 Summary

Coming back to our first research question, how far can we provide empirical evidence that the measurement instruments developed within the TEDS-M research program have predictive validity for teaching mathematics at high quality, we could show that especially the content-related knowledge parts of teachers competence together with noticing as situation-specific competence facets influence the instructional quality of their lessons. Especially the identification of different teacher profiles explained the relation between the knowledge part of teacher competence and the situation-specific skills with instructional quality under a content-related and a pedagogical perspective (e.g. related with adaptivity).

Concerning the second research question of the added value of the video-based measured situation-specific competence facets our study reveals lower correlations between teachers' professional noticing and domain-specific dimensions of instructional quality as expected. Especially the comparative study between Chinese and German in-service teachers showed culture-specific strengths and weaknesses of both teacher groups.

5 Outlook

Central lessons learned from the TEDS-M research program and specifically TEDS-Validate are as follow: We consider to highlight the empirical validation of the multidimensionality of teacher competence and, therefore, derive the need for evaluating teacher competences with different assessment methods focusing on different sub-facets of teacher competence. Especially professional noticing consisting of the situated competence facets play an important role for providing instructional quality from a general pedagogical perspective as well as from a content-related perspective. Thereby, these situated competence facets have the potential to integrate subject-based professional knowledge with general pedagogical professional knowledge via teaching incidents and classroom activities. The central lesson learned leads to the following research focus in the coming transfer phase of TEDS-Validate.

Initial teacher education requires elaborated and carefully designed school-based practical opportunities to learn – often related to as various types of school-based practicum. These learning opportunities intend to enable pre-service teachers to relate their professional knowledge gained during the academic part of teacher education with practical teaching situations and the associated situation-specific requirements. The application of theoretical knowledge in practical situations, for example the implementation of teaching plans and the handling of students' reactions in specific learning situations, are general objectives of practical phases of teacher education. However, the extent to which professional teaching skills can be promoted in such practical phases during university studies is still a largely open question. In their survey of the state of research in the German-speaking discussion, König et al. (2018) noted a clear lack of empirical studies that could prove the effects of extended practice in initial teacher education. An international review also concluded that the state of research on the effectiveness of extended practicum phases in teacher education can be described as very narrow (Lawson et al. 2015).

To address this research gap the planned and already approved transfer study of TEDS-Validate will transfer the conceptualizations and evaluation instruments

from TEDS-Validate into the practicum phase within the master studies of initial teacher education. The video-based tests for professional noticing including perception, interpretation and decision-making developed for in-service mathematics teachers and validated in the current project will be used with samples of pre-service teachers in the master phase of six German universities from several federal states. As the project partners will benefit through their involvement and support of data collection and developing a shared understanding of findings, these planned transfer activities can therefore be expected to provide a clear added value in terms of dissemination of the instruments developed and the use of the findings gained with the help of these instruments for the future design of practical learning opportunities in the field of initial teacher education. Moreover, the transfer of major insights into teacher competence as provided by TEDS-Validate into higher education will clearly contribute to important discussions on reforming teacher education such as the ongoing debate about an adequate ratio of theory and practice or about making concrete what kind of competencies pre-service teachers should elaborate on during their professionalization process and career development.

To sum up, the TEDS-M research program provides a broad overview on the evaluation of initial teacher education and professional development of teachers, its efficiency and necessity for further developments.

References

- Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9, 469–520.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47, 133–180.
- Benecke, K. (2018). Messung von Unterrichtsqualität durch Unterrichtsbeobachtungen – eine Studie zum Vergleich von Live- und Video-Rating. In R. Biehler et al. (Eds.), *Beiträge zum Mathematikunterricht 2018* (pp. 2063–2064). Münster: WTM-Verlag.
- Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society*, 24(3), 200–212.
- Blömeke, S. (2017). Modelling teachers' professional competence as a multi-dimensional construct. S. Guerriero (Ed.), *Pedagogical Knowledge and the Changing Nature of the Teaching Profession* (pp. 119–135). Paris: OECD.
- Blömeke, S. & Delany, S. (2012). Assessment of teacher knowledge across countries: A review of the state of research. *ZDM Mathematics Education*, 44, 223–247.
- Blömeke, S., Gustafsson, J.-E. & Shavelson, R. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13.

- Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.). (2010a). *TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich*. Münster: Waxmann.
- Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.). (2010b). *TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte im internationalen Vergleich*. Münster: Waxmann.
- Blömeke, S., Bremerich-Vos, A., Haudeck, H., Kaiser, G., Lehmann, R., Nold, G., Schwip-pert, K., & Willenberg, H. (Eds.). (2011). *Kompetenzen von Lehramtsstudierenden in gering strukturierten Domänen. Erste Ergebnisse aus TEDS-LT*. Münster: Waxmann.
- Blömeke, S., Hsieh, F.-J., Kaiser, G., & Schmidt, W.H. (Eds.). (2014). *International Perspectives on Teacher Knowledge, Beliefs and Opportunities to Learn*. Dordrecht: Springer.
- Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M. & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf – Fachbezogenes Wissen als Voraussetzung für Wahrnehmung, Interpretation und Handeln im Unterricht. *Zeitschrift für Erziehungswissenschaft*, 17 (3), 509–542.
- Blömeke, S., Kaiser, G., König, J., & Jentsch, A. (accepted, 2020). Profiles of mathematics teachers' knowledge, beliefs, and skills, and their relation to instructional quality. *ZDM Mathematics Education*, 52(3).
- BMBF (2014) = Bundesministerium für Bildung und Forschung. (2014). *Bekanntmachung des Bundesministeriums für Bildung und Forschung von Richtlinien zur Förderung der „Qualitätsoffensive Lehrerbildung“*. Berlin: BMBF.
- Bruckmaier, G., Krauss, S., Blum, W., & Leiss, D. (2016). Measuring mathematical teachers' professional competence by using video clips (COACTIV video). *ZDM Mathematics Education*, 48 (1), 111–124.
- Buchholtz, N., Scheiner, T., Döhrmann, M., Suhl, U., Kaiser, G., & Blömeke, S. (2016). *TEDS-shortM. Kurzfassung der mathematischen und mathematikdidaktischen Testinstrumente aus TEDS-M, TEDSLT und TEDS-Telekom*. Hamburg: Universität Hamburg.
- Cochran-Smith, M., & Zeichner, K.M. (2005). *Studying teacher education: The report of the AERA Panel on Research and Teacher Education*. Mahwah, NJ: Erlbaum.
- Cronbach, L. J., Gleser, G., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis*, 23(1), 57–77.
- Embretson, S. (1983). Construct validity. Construct representation versus the nomothetic span. *Psychological Bulletin*, 93 (1), 179–197.
- Flores, M. A. (2016). Teacher Education Curriculum. In J. Loughran, & M.L. Hamilton (Eds.), *International Handbook of Teacher Education* (pp. 187–230). Dordrecht: Springer.
- Großschedl, J., Harms, U., Kleickmann; T. and Glowinski, I. (2015) 'Preservice Biology Teachers' Professional Knowledge: Structure and Learning Opportunities'. *Journal of Science Teacher Education*, 26(3), 291–318.
- Guerriero, S. (Eds.). (2017). *Pedagogical Knowledge and the Changing Nature of the Teaching Profession*. Paris: OECD.
- Hattie, J. (2009). *Visible Learning: a synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

- Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze-Velber: Klett/Kallmeyer.
- Hill, H. C., Rowan, B. & Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal*, 42, 371–406.
- Jentsch, A., Schlesinger, L., Heinrichs, H., Kaiser, G., König, J., & Blömeke, S. (under review). Unterrichtsqualität unter einer mathematikdidaktischen Perspektive – Konzeptualisierung, Messung und Validierung. *Journal für Mathematik-Didaktik*.
- Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M., et al. (2006). *Klassifikationsschema für Mathematikaufgaben: Dokumentation der Aufgabekategorisierung im COACTIV-Projekt*. Berlin: Max-Planck-Institut für Bildungsforschung.
- Kaiser, G., Blömeke, S., König, J., Busse, A., Döhrmann, M. & Hoth, J. (2017). Professional competencies of (prospective) mathematics teachers – cognitive versus situated approaches. *Educational Studies in Mathematics*, 94 (2), 161–182.
- Kaiser, G., Busse, A., Hoth, J., König, J. & Blömeke, S. (2015). About the Complexities of Video-Based Assessments: Theoretical and Methodological Approaches to Overcoming Shortcomings of Research on Teachers' Competence. *International Journal of Science and Mathematics Education*, 13(2), 369–387.
- Kaiser, G., & König, J. (2019). Competence Measurement in (Mathematics) Teacher Education and Beyond: Implications for Policy. *Higher Education Policy*, 32, DOI: 10.1057/s41307-019-019-00139-z.
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R. & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, 49, 568–589.
- Klemenz, S. & König, J. (2019). Modellierung von Kompetenzniveaus im pädagogischen Wissen bei angehenden Lehrkräften: Zur kriterialen Beschreibung von Lernergebnissen der fächerübergreifenden Lehramtsausbildung. *Zeitschrift für Pädagogik*, 65 (3), 355–377.
- Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, 52(6), 876–903.
- König, J. (2009). Zur Bildung von Kompetenzniveaus im Pädagogischen Wissen von Lehramtsstudierenden: Terminologie und Komplexität kognitiver Bearbeitungsprozesse als Anforderungsmerkmale von Testaufgaben?. *Lehrerbildung auf dem Prüfstand*, 2(2), 244–262.
- König, J. (2019). Pedagogical Knowledge in Teacher Education. In M. A. Peters (Eds.), *Encyclopedia of Teacher Education*. Dordrecht: Springer.
- König, J. (2014). *Designing an International Instrument to Assess Teachers' General Pedagogical Knowledge (GPK): Review of Studies, Considerations, and Recommendations*. Technical paper prepared for the OECD Innovative Teaching for Effective Learning (ITEL) – Phase II Project: A Survey to Profile the Pedagogical Knowledge in the Teaching Profession (ITEL Teacher Knowledge Survey). OECD: Paris.
- König, J. (2015). Measuring Classroom Management Expertise (CME) of Teachers: A Video-Based Assessment Approach and Statistical Results. *Cogent Education*, 2 (1), 991178.

- König, J. & Blömeke, S. (2010). *Pädagogisches Unterrichtswissen (PUW). Dokumentation der Kurzfassung des TEDS-M-Testinstruments zur Kompetenzmessung in der ersten Phase der Lehrerbildung*. Berlin: Humboldt-Universität.
- König, J. & Blömeke, S. (2013). Preparing Teachers of Mathematics in Germany. In J. Schulle, L. Ingvarson & R. Holdgreve-Resendez (Eds.), *TEDS-M Encyclopaedia. A Guide to Teacher Education Context, Structure and Quality Assurance in 17 Countries. Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)* (pp. 100–115). Amsterdam: IEA.
- König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A., & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education*, 38, 76–88.
- König, J., Blömeke, S., Paine, L., Schmidt, B. & Hsieh, F.-J. (2011). General Pedagogical Knowledge of Future Middle School Teachers. On the Complex Ecology of Teacher Education in the United States, Germany, and Taiwan. *Journal of Teacher Education*, 62 (2), 188–201.
- König, J., Kaiser, G., Blömeke, S., Jentsch, A., Schlesinger, L., Nehls, C., & Suhl, U. (under review). Teaching and learning in the lower secondary mathematics classroom: Analyzing the links between pedagogical competence, instructional quality, and student achievement.
- König, J., & Kramer, C. (2016). Teacher professional knowledge and classroom management: On the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). *ZDM Mathematics Education*, 48 (1), 139–151.
- König, J., Lammerding, S., Nold, G., Rohde, A., Strauß, S. & Tachtsoglou, S. (2016). Teachers' Professional Knowledge for Teaching English as a Foreign Language: Assessing the Outcomes of Teacher Education. *Journal of Teacher Education*, 67 (4), 320–337.
- König, J., & Pflanzl, B. (2016). Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. *European Journal of Teacher Education*, 39 (4), 419–436.
- König, J., Rothland, M., & Schaper, N. (Eds.). (2018). *Learning to Practice, Learning to Reflect? Ergebnisse aus der Längsschnittstudie LiP zur Nutzung und Wirkung des Praxissemesters in der Lehrerbildung*. Wiesbaden: Springer.
- Kraimer, K., & Llinares, S. (2010). Mathematics teacher education, in P. Peterson, E. Baker, & B. McGaw (Eds.). *International Encyclopedia of Education* (pp. 702–705). Oxford: Elsevier.
- Krauss, S., Lindl, A., Schilcher, A., Fricke, M., Göhring, A. & Hofmann, B. (Eds.). (2017). *FALKO: Fachspezifische Lehrerkompetenzen: Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik*. Münster: Waxmann.
- Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms CO-ACTIV*. Münster: Waxmann.
- Lawson, T., Cakmak, M., Gündüz, M., & Busher, H. (2015). Research on teaching practicum – a systematic review. *European Journal of Teacher Education*, 38(3), 392–407.
- Nehls, C., König, J., Kaiser, G., & Blömeke, S. (2020). Profiles of teachers' general pedagogical knowledge: Nature, causes and effects on beliefs and instructional quality. *ZDM Mathematics Education*, 52(3). <https://doi.org/10.1007/s111858-010-01102-3>.

- Nehls, C., König, J., Kaiser, G., Klemenz, S., Ross, N., & Blömeke, S. (under review). Pädagogisches Wissen von berufstätigen Mathematiklehrkräften – Validierung der Konstruktrepräsentation im TEDS-M-Instrument. *Diagnostica*.
- Palardy, G. J. & Rumberger, R.W. (2008). Teacher Effectiveness in First Grade: The Importance of Background Qualifications, Attitudes, and Instructional Practices for Student Learning. *Educational Evaluation and Policy Analysis*, 30, 111–140.
- Pankow, L., Kaiser, G., Busse, A., König, J., Blömeke, S., Hoth, J., & Döhrmann, M. (2016). Early career teachers' ability to focus on typical students errors in relation to the complexity of a mathematical topic. *ZDM Mathematics Education*, 48(1–2), 55–67.
- Ross, N., & Kaiser, G. (2018). Klassifikation von Mathematikaufgaben zur Untersuchung mathematisch-kognitiver Aspekte von Schülerleistungstests und von Unterrichtsqualität. In R. Biehler et al. (Eds.), *Beiträge zum Mathematikunterricht 2018* (pp. 1519–1522). Münster: WTM-Verlag.
- Rowland, T., & Ruthven, K. (Eds.). (2011). *Mathematical knowledge in teaching*. Dordrecht: Springer.
- Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM Mathematics Education*, 50 (3), 475–490.
- Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Educational Research Journal*, 51(4), 739–771.
- Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (Eds.). (2011). *Mathematics Teacher Noticing. Seeing Through Teachers' Eyes*. New York: Routledge.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Research*, 57, 1–22.
- Stahnke, R., Schueler, S., & Roesken-Winter, B. (2016). Teachers' perception, interpretation, and decision-making: a systematic review of empirical mathematics education research. *ZDM Mathematics Education*, 48(1–2), 1–27.
- Stigler, J. W., & Miller, K. F. (2018). Expertise and Expert Performance in Teaching. In A. Ericsson, R.R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 431–452). Cambridge: Cambridge University Press.
- Tatto, M. T., & Senk, S. (2011). The mathematics education of future primary and secondary teachers: Methods and findings from the Teacher Education and Development Study in Mathematics. *Journal of Teacher Education*, 62(2), 121–137.
- Voss, T., Kunter, M., Seiz, J., Hoehne, V., & Baumert, J. (2014). Die Bedeutung des pädagogisch-psychologischen Wissens von angehenden Lehrkräften für die Unterrichtsqualität. *Zeitschrift für Pädagogik*, 60(2), 184–201.
- Weinert, F. E. (2001). Concept of Competence: A Conceptual Clarification. In D. S. Rychenij & L.H. Salganik (Eds.), *Defining and Selecting Key Competencies* (pp. 45–66). Göttingen: Hogrefe.
- Yang, X., Kaiser, G., König, J., & Blömeke, S. (2018). Measuring Chinese Teacher Professional Competence: Adapting and Validating a German Framework in China. *Journal of Curriculum Studies*, 50 (5), 638–653.
- Yang, X., Kaiser, G., König, J. & Blömeke, S. (2019). Professional Noticing of Mathematics Teachers: a Comparative Study Between Germany and China. *International Journal of Science and Mathematics Education*, 17(5), 943–963.



2.3 Planning Competence of Pre-Service German Language Teachers

Conceptualization, Measurement, and Validation

König, J., Bremerich-Vos, A., Buchholtz, C., Fladung, I., and
Glutsch, N.

Abstract

The research project PlanvoLL-D carried out from 2016 to 2019 aimed at providing new insights into the modelling and measurement of teacher planning competence by analyzing pre-service teachers' written plans of demonstration lessons in a standardized way. In this chapter, we outline the theoretical framework of the PlanvoLL-D project and its study design. We provide insights into the conceptualization and measurement of pre-service language teacher planning competence and give a summary of the project's major findings. Finally, we discuss implications for teacher education and give an outlook for further research.

Keywords

Language teaching, lesson planning, pre-service teachers, planning competence, teacher education

1 Introduction

Planning lessons is part of the core tasks for the professional teacher. Teacher education programs in Germany and in many other countries worldwide intend to train pre-service teachers planning a single lesson or a unit of lessons, so that they can master the daily work of lesson planning when entering teaching (e.g., European Commission 2013; König et al. 2017a). Initial teacher education provides pedagogical, subject-specific and practical learning opportunities that also relate to the area of lesson planning. Numerous books exist offering theories or practical guidelines of lesson planning to pre-service teachers (e.g., John 2006; Scholl 2018) and teacher-certification procedures such as the state examination in the German induction phase (“Vorbereitungsdienst” or “Referendariat”) usually require planning demonstration lessons (e.g., König and Blömeke 2010; Kärner et al. 2019; Pecheone and Chung 2007; Strietholt and Terhart 2009). However, empirical research on lesson planning as a skill for pre-service teachers and the development of such skills during initial teacher education is scarce (König et al. 2015; Cochran-Smith and Villegas 2016).

Against this background, between 2016 and 2019, the research project PlanvoLL-D was carried out. The PlanvoLL-D project is entitled “The Role of Professional Knowledge of Pre-Service German Teachers in their Lesson Planning” [“Die Bedeutung des professionellen Wissens angehender Deutschlehrkräfte für ihre Planung von Unterricht“]. PlanvoLL-D aimed at providing new insights into the modelling and measurement of teacher planning competence by analyzing pre-service teachers’ written plans of demonstration lessons in a standardized way. In this chapter, we outline the theoretical framework of the PlanvoLL-D project and its study design. We provide insights into the conceptualization and measurement of per-service teacher planning competence and give a summary of the project’s major findings. Finally, we discuss implications for teacher education and give an outlook for further research.

2 Theoretical Framework

2.1 Investigation into the Field of Lesson Planning

Although planning lessons belongs to the daily work of teachers (Hardwig and Mußmann 2018), empirical research in this area is scarce. Hardly any approaches exist that measure planning skills in a standardized way (König 2019). Early studies in lesson planning investigated specific aspects, for example, which planning

component a teacher pays attention to and the order in which a teacher works on such planning components (e.g. Clark and Peterson 1986; Zahorik 1975; Taylor 1970; Hill et al. 1983). For example, an in-service teacher survey conducted by Taylor (1970) provided evidence that during lesson planning teachers gave priority to student needs, learning content, goals, and methods. Teachers first started with thinking of the teaching context (comprising materials, resources), then with involving the specific student needs and learning dispositions, and finally thinking of curricular alignment. Hill, Yinger, and Robbins (1983) showed in a similar way that after selecting appropriate materials, teachers gave priority to planning decisions of how they can arrange these materials in the classroom so that their students use them as activities. Lesson planning procedures of this kind can be described as a problem-solving process (e.g., Yinger 1977; Bromme 1981), highlighting the decisions teachers make on the basis of available information during pre-active teaching (Shavelson and Borko 1979) and as part of their reflection on action (Parsons et al. 2018). Relevant skills can be considered as part of teacher competence and therefore should be an object of empirical investigation (König et al. 2015).

The research on the measurement of teacher competence has significantly increased over the last decade (Baumert 2016; Kaiser and König 2019; see also chapters by Kuhn et al.; Vogelsang et al.; Lemmrich et al. in this volume). Various research groups have developed standardized test instruments assessing teacher knowledge following the well-known classification by Shulman (1987): content knowledge (CK), pedagogical content knowledge (PCK), and general pedagogical knowledge (GPK). Moreover, recent research has started to assess teachers' situation-specific skills in the area of professional vision (Kaiser et al. 2015; Kaiser and König in this volume). However, to the best of our knowledge, situation-specific skills in the area of lesson planning has not been an object of investigation in the modelling and measuring of teacher competence.

As a consequence, assumptions about processes of lesson planning seem hardly be supported by empirical evidence. For example, it remains an open question, whether teachers presumably make use of their specific knowledge in the subject area, subject-specific pedagogy, and general pedagogy and relate it to the specific planning situation that is predominantly determined by factors such as characteristics of the learning group, specific curricular goals, and the classroom context. At least some evidence exists that didactical models – being predominantly prescriptive rather than evidence-based – are not necessarily applied by in-service teachers, although they may be given high priority in initial teacher education programs (John 2006; Scholl 2018).

Against this background of investigations into the field of lesson planning, the PlanvoLL-D project aimed at a valid measurement of pre-service teachers' skills in the area of lesson planning. However, due to the complexity of lesson planning as an object of investigation, only a particular, but highly relevant aspect was examined: Our focus is on the construct of *pedagogical adaptivity*, that is, the ways in which the assignments of the lesson matches the cognitive level of the learning group (König et al. 2015; for a more detailed description, see König et al. 2019).

2.2 Learning Dispositions of Students

The research on teacher expertise has proliferated important insights into the lesson planning of expert and novice teachers (Stigler and Miller 2018). Expert teachers plan their lessons in a process-driven way and are capable to consider several planning elements simultaneously. They rigorously relate the learning dispositions (e.g., domain-specific knowledge) of their students and the learning tasks chosen for the lesson to each other (Berliner 2004; Borko and Livingston 1989; Housner and Griffey 1985; Smith and Strahan 2001; Westermann 1991). Expert teachers are clearly aware of their students and are committed to involve student needs into their planning process. They perceive student learning dispositions as a key element of their teaching and know how to integrate diagnostic information from their students specifically into their lesson planning (Putnam 1987). When making decisions during the planning process they manage to integrate their conceptual and situation-specific knowledge (De Jong and Ferguson-Hessler 1996). Important aspects of the planning situation are identified and are progressively merged with teaching and learning activities (Ericsson 1996; Leinhardt and Greeno 1986; Schoenfeld 1998).

2.3 Planning Learning Tasks

When planning a lesson, teachers select and create learning tasks as part of student activities in the classroom. Here, teachers are also able to integrate a range of further decisions, for example regarding the selection of content and the specification of objectives that are part of the lesson (Bromme 1981; Kang 2017; Shavelson and Stern 1981; Yinger 1977; Zahorik 1975). Learning tasks usually reflect the objectives of a lesson, since they refer to what students should learn, what knowledge students should acquire, or which competencies students should elaborate. Classification systems or taxonomies support the analysis of learning tasks in relation to

specific cognitive and motivational requirements (Anderson and Krathwohl 2001; Commons et al. 1998; De Jong and Ferguson-Hessler 1996; Johnson 2003). At the same time, complex learning tasks can cover a range of difficulty levels in different dimensions, allowing an alignment to existing student dispositions and needs in a differentiated way. They therefore serve to implement teaching strategies of differentiated instruction and support teachers when they account for the existing knowledge of learners and guide learners into their “zone of proximal development” (Vygotsky 1978, p. 84). Learning tasks can therefore be regarded as an important instrument of adaptive teaching (Corno 2008; Parsons et al. 2018). The way a teacher deals with learning tasks during lesson planning might provide insight into his or her pedagogical adaptivity.

In the PlanvoLL-D project, we specifically look at those learning tasks students are required to work on during the lesson’s main activity phase. These tasks represent the work that the teacher instructs his or her students to engage in. The work is expected to trigger in students cognitive activation and information processing (Neubrand et al. 2013). Students usually work on such tasks individually, in pairs, or sometimes in groups. Usually these tasks can be clearly identified in written lesson plans (König et al. 2015), as they emerge from the relevant lesson material (e.g., a worksheet or a number of differentiated worksheets) that guides student work.

2.4 Lesson Planning as Part of Teacher Competence

Due to little investigation into this field, empirical evidence on how teachers plan their lessons is fairly limited (Bromme 1981; Jacobs et al. 2008). Although some surveys or qualitative studies exist that provide relevant descriptive scientific knowledge, to generate explanatory knowledge, approaches that directly assess teacher skills in the area of lesson planning are necessary. An exception is the *Performance Assessment for California Teachers* (PACT): It requires pre-service teachers to complete several components related to planning lessons, teaching, assessing students, and reflecting on teaching, where they are asked to submit an outline for three to five lessons they are going to teach. The performance ratings are based on coding schemes with a 4-point continuum. For the task “planning”, five guiding questions are used by the raters who have to score the quality of the instructional design (Pecheone and Chung 2007, p. 27). These are related to how the instructional design provides students to have access to the curriculum, how the curriculum is addressed in a coherent and balanced way, how the students’ interest and needs are reflected and addressed, and how well learning goals, in-

struction, and assessments are aligned. PACT provides important insights into a teacher performance assessment that is very close to typical tasks teachers have to master. In a more recent analysis on predictive validity, Darling-Hammond et al. (2013) showed that the PACT overall scale, as well as subscales such as planning, can significantly predict student achievement. However, since information on the scaling procedure is limited (Pechone and Chung 2007), we conclude that research on measuring and modelling lesson planning as part of teacher professional competence can still be regarded as a research desideratum.

2.5 Modelling of Planning Competence in PlanvoLL-D

Planning a lesson is dependent on the context in general (John 2006; Mutton et al. 2011). Therefore, it is important to account for the situation of planning that is determined by characteristics of the learning group, curricular goals, or the institutional setting. Following the model of “competence as a continuum” as suggested by Blömeke et al. (2015), we make the following assumptions that underlie our investigation of pedagogical adaptivity. First, in the project PlanvoLL-D, we define pedagogical adaptivity as a situation-specific teacher skill. Teacher professional knowledge as investigated as “cognitive disposition” (Blömeke et al. 2015) by previous studies should be a relevant antecedent of such a situation-specific skill. We consider situation-specific lesson planning skills as being more proximal to actual performance in class than teacher knowledge. Figure 1 illustrates this idea. It serves as a heuristic to locate the constructs of professional competence in the PlanvoLL-D project in an overall model.

In PlanvoLL-D, pre-service teacher professional knowledge needed for lesson planning was investigated using standardized tests measuring their content knowledge (CK), pedagogical content knowledge (PCK), and general pedagogical knowledge (GPK). By contrast, pedagogical adaptivity as a skill was investigated using the data of authentic lesson plans that de facto were enacted as demonstration lessons during the second phase of initial teacher education, i.e., the induction phase in Germany (“Vorbereitungsdienst” or “Referendariat”). The nature of pedagogical adaptivity as a situation-specific skill therefore is different from CK, PCK, or GPK as measured using the paper-pencil approach. Instructional practice as an indicator of classroom performance was captured using self-reports of pre-service teachers they were asked to provide for the specific lesson after performing that lesson (for further information on the instruments, see Section 3.4).

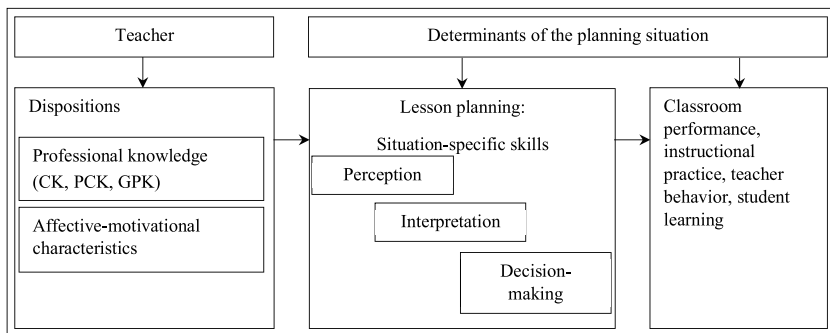


Figure 1 Heuristic of modelling lesson planning as part of teacher competence in PlanvoLL-D (following Blömeke et al. 2015, p. 7)

2.6 Lesson Planning as Part of Teacher Education

Teacher education programs intend future teachers to learn how to plan lessons. Corresponding opportunities to learn are provided by teacher education institutions in many countries. While courses in the academic setting often primarily aim at the acquisition of theoretical knowledge, in-school opportunities to learn give future teachers the chance to connect their knowledge to practical situations in the classroom (König et al. 2017a). Lesson planning might be a particularly complex challenge for novice teachers, as they are required to link their professional knowledge to the concrete learning group they are going to teach (John 2006). An analysis of how pre-service teachers, during induction, relate their lesson planning to previously acquired theoretical knowledge might, therefore, be a valuable contribution to the teacher education theory-practice discourse.

2.7 Context of the PlanvoLL-D Project

The German teacher education system has a consecutive structure with two separate phases, a theoretical at university and a practical at small teacher training institutions operated by state governments (König and Blömeke 2013). Pre-service teachers finish their first phase at university with a master of education nowadays, requiring coursework that emphasizes the acquisition of theoretical knowledge in the teaching subjects, both subject-specific as well as general pedagogical knowledge (König et al. 2017a). By contrast, most practical learning opportunities are

then provided in the 1.5-year second phase. This phase serves as induction for the pre-service teachers who then work part-time at schools and attend courses in general pedagogy and subject-related pedagogy. They are assessed by their teacher educators and mentored by one or two teachers at school. Lesson performance is usually based on a written lesson plan comprising detailed information about a large number of planning aspects such as objectives, teaching methods, the learning group, activities, time schedule, and embedding the lesson into the larger teaching unit. For this, pre-service teachers are required to have observed or even taught the learning group in advance, so that they are familiar with the students and had the opportunity to learn about the students' prior knowledge and motivation. Pre-service teachers are required to give demonstration lessons and to submit the relevant written plan at regular intervals over the duration of the second phase. This phase ends with a state examination consisting of a practical part including at least two lessons performed in two different subjects.

3 Study Design

3.1 Research Questions

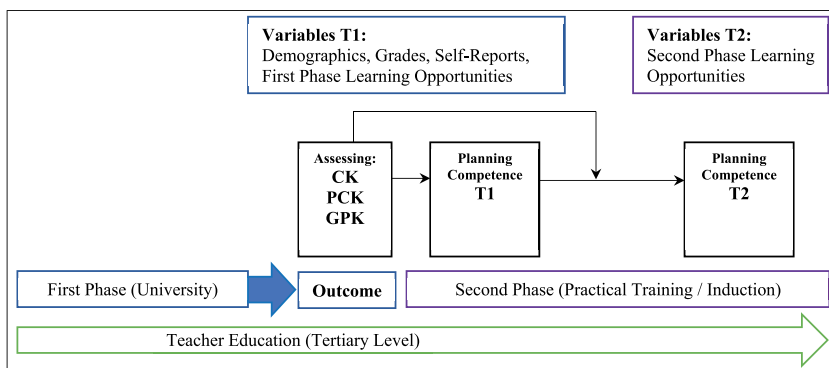
The PlanvoLL-D project aimed at the investigation of the planning competence of pre-service teachers for German during induction (“Vorbereitungsdienst” or “Referendariat”). We focused on two major research questions:

1. Is it possible to differentiate the planning competence of pre-service language teachers into generic and subject-specific lesson-planning skills?
2. What are potential factors influencing the pre-service teachers' planning competence, such as specific aspects of their learning opportunities during the second phase of initial teacher education (i.e., the induction phase) or the professional knowledge they had acquired during the first phase of initial teacher education at university?

3.2 Research Model

To examine our research questions, we decided to carry out the PlanvoLL-D project in the second phase of initial teacher education that serves as induction for pre-service teachers. In our research model (Figure 2), professional knowledge of pre-service teachers is defined as an outcome of the first phase of teacher prepa-

ration at university. That knowledge was tested at the beginning of the second phase, using the tests measuring CK in linguistics and literature, PCK, and GPK (Bremerich-Vos et al. 2019; König et al. 2011; König and Bremerich-Vos 2019). Moreover, pre-service teachers' planning competence was measured at two time points: We asked pre-service teachers to submit the written plans of the lessons they demonstrated at the very start and the very end of their practical training. In the latter case, the demonstration lesson was part of the certification procedure (state examination). This approach enables us to examine planning competence at two time points, and also to analyze potential influencing factors such as professional knowledge and learning opportunities. Pre-service teachers were asked to report on the learning opportunities they had been exposed to, using different scale inventories that relate to the first phase at university and the induction phase (Glutsch et al. 2019; König et al. 2017). Table 1 provides an overview of the instruments used in the PlanvoLL-D project.



Abbreviations: CK – German Content Knowledge, PCK – Pedagogical Content Knowledge, GPK – General Pedagogical Knowledge; T1 – Time Point 1, T2 – Time Point 2

Figure 2 Research model of the PlanvoLL-D project

Table 1 Overview of the instruments used in the PlanvoLL-D project

Name of instrument	Type of instrument	Reference for further reading
CK – German Content Knowledge	Paper-Pencil Test	Bremerich-Vos, König, & Fladung 2019; König & Bremerich-Vos 2019
PCK – Pedgogical Content Knowledge	Paper-Pencil Test	Bremerich-Vos, König, & Fladung 2019; König & Bremerich-Vos 2019
GPK – General Pedgogical Knowledge	Paper-Pencil Test	König & Bremerich-Vos 2019
Planning Competence	Sample Work Analysis	König, J., Bremerich-Vos, A., Buchholtz, C., Fladung, I., & Glutsch, G. 2019
Opportunities to Learn	Questionnaire	Glutsch, N., Bremerich-Vos, A., Buchholtz, C., König, J., Fladung, I., Lammerding, S., Strauß, S., & Schleiffer, C. 2019; König et al. 2017a

3.3 Sample

Data was collected in two federal states, North Rhine-Westphalia and Berlin. Our target group was defined as pre-service secondary teachers for German who had entered the second phase of teacher education in spring 2016. Two teaching types were included in North Rhine-Westphalia: Pre-service teachers attending a teacher education program that would qualify them as lower secondary teachers only (*Haupt-/Real-/Gesamtschule*) or as lower and upper secondary teachers (*Gymnasium/Gesamtschule*). In Berlin, the corresponding teacher education program focused on was a comprehensive teacher education program that would qualify pre-service teachers as lower and upper secondary teachers (*Integrierte Sekundarschule/Gymnasium*). A random sample of training units was drawn for the lower and upper secondary teacher education program in North Rhine-Westphalia, whereas, due to smaller populations, a census was applied for the other two programs. Participation rate on the level of training units was good (92% in North Rhine-Westphalia) or at least acceptable (70% in Berlin). Within these training units, all pre-service teachers were included in the survey. Participation rate on the individual level was good (91% in Berlin) or at least acceptable (68% in North Rhine-Westphalia). The sample of this first time point (T1) consists of 378 pre-service teachers (Figure 3).

Research assistants of the project team administered a paper-pencil questionnaire that the pre-service teachers completed under observation. This questionnaire included the standardized test to assess pre-service teachers' PCK and GPK as well as other instruments (e.g., on learning opportunities at university; for more details, see Glutsch et al. 2019; König et al. 2017a). Data collection was continued online, also comprising a third test to assess pre-service teachers' CK (Figure 2). After the survey, pre-service teachers were asked to submit a copy of the written plan of their first demonstration lesson and to complete a short questionnaire related to the execution of that lesson (Figure 3). With 172 plans and questionnaires submitted that finally could be linked with the previous survey data of 378 pre-service teachers, participation rate was moderate (46 %); however, a drop out analysis did not show sample bias.

About 1.5 years later, pre-service teachers were re-examined, resulting in a second time point (T2) with 138 pre-service teachers who submitted a copy of the written plan from their last demonstration lesson (state examination) and again completed a short questionnaire (Figure 3). They were asked finally to participate in another online survey in which they had to report on their second phase learning opportunities. 130 pre-service teachers participated in this final survey (Glutsch et al. 2019). Lesson plans from first and second time point allow us to analyze a panel of 116 pre-service teachers.

Additional data collection was carried out in North Rhine-Westphalia only between the two time points (Figure 3): Pre-Service teachers were another time asked to submit a copy of an intermediate written plan of a demonstration lesson, but this was also linked to the requirement to apply a short questionnaire on instructional quality as rated by their students. 27 pre-service teachers and 564 school students participated in this additional data collection component (for the first findings, see König and Bremerich-Vos 2019).

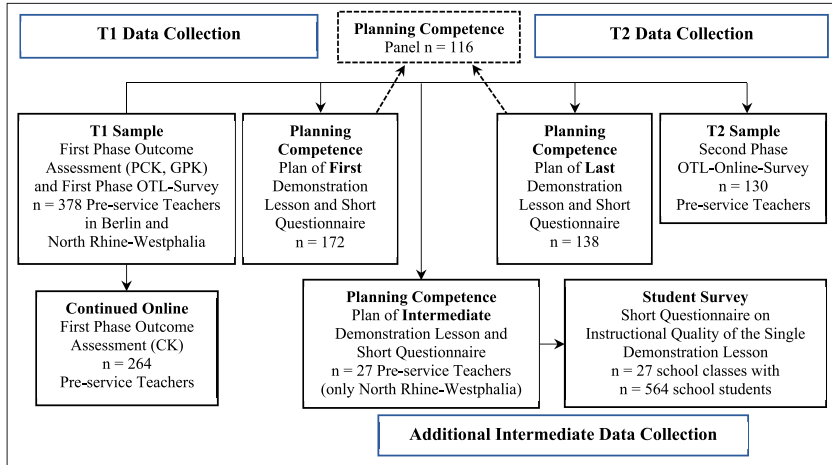


Figure 3 Data collection components of the PlanvoLL-D project

3.4 Instrument: Pedagogical Adaptivity in Written Plans of Demonstration Lessons

PlanvoLL-D builds on a previous study on planning competence: „*Planungskompetenz von Lehrerinnen und Lehrern*“ (PlanvoLL) (König et al. 2015; Buchholtz and König 2015). In the PlanvoLL project, we provided a first approach on measuring and modelling lesson planning competence, by analyzing written plans of demonstration lessons (“Lehrproben”) in a standardized way. Applying this methodological approach to the planning process implies that the pre-service teachers are required to relate their knowledge to a concrete, real learning group. It thus differs from standardized knowledge tests that require pre-service teachers to respond to test items capturing aspects of lesson planning in a de-contextualized way. However, with the innovative approach developed in the PlanvoLL project, pedagogical adaptivity was captured irrespective of the subject only. Since the purpose of lesson planning is always related to a specific subject, the domain-specific modelling and measuring of planning competence remains a research gap.

To investigate pedagogical adaptivity of pre-service teachers in their written lesson plans as a situation-specific skill, the written plans of demonstration lessons were analyzed and indicators created on the basis of the existing coding system developed in PlanvoLL. The coding system differentiates indicators into four components (Figure 4): On a descriptive level, written plans are analyzed whether the

learning group is described (component 1) and whether descriptive information is given for the learning task that primarily governs students' activities during the lesson (component 2). On an analytical level, plans are analyzed whether the descriptions given by a pre-service teacher for his or her specific learning group and the learning task or tasks planned are logically and pedagogically consistent (component 3). This application of the given descriptions to the specific situation comprises the examination whether the learning task is adapted to the cognitive level of students following the "zone of proximal development" (Vygotsky 1978, p. 84). Therefore, it is necessary that the lesson plan contains an outline that shows how the task (or even a differentiated set of tasks) given to the students connects with what the students (or groups of students) have learned so far, for instance, in a preceding lesson of the unit that contextualizes the plan of the demonstration lesson. The connection between tasks and prior knowledge of students needs to be addressed by the pre-service teacher and he or she should relate this connection to the situation of the particular lesson. Finally, plans are analyzed to determine whether such adaptive teaching is linked to other important elements of planning such as the connection to learning goals (component 4). As an innovative element of the PlanvoLL-D project, analogous to the generic indicators, subject-specific indicators were created to measure pre-service teachers' elaboration on content-specific planning (Figure 4). For example, a statement like "The learning group (...) consists of 29 students, twelve girls and 17 boys." appears as a quite short description of the learning group in one lesson plan without a content-specific planning. In another lesson plan, by contrast, a pre-service teacher shows his ability to detail the subject-specific characteristics of his learning group: „(...) The students are familiar with writing poems of different types according to criteria from the previous grade. Using a sample poem for writing has already been trained in the course of the unit and is also familiar from the previous grade." (for more coding examples, see König et al. 2019).

Altogether, the coding scheme in Figure 4 consists of 23 indicators (13 generic and 10 subject-specific). Since frequency distribution showed that a subject-specific indicator was fulfilled less frequently than the corresponding generic indicator, we constructed partial-credit items (Masters 2016) with scoring category 1 (generic indicator fulfilled) and 2 (subject-specific indicator fulfilled). Therefore, our scaling analysis comprises 13 items (10 partial-credit, three dichotomous) making up a reliable scale (*EAP* reliability .79, comparable with *Cronbach's alpha*). Missing values (code 9) were made explicit in the scaling model and adequately accounted for in the IRT scaling analysis (for more details on the IRT scaling procedure, see König et al. 2019).

Component	Sample generic indicator	Number of generic indicators (items)	Number of subject-specific indicators (items)
(1) Description of situation-specific factors	The teacher describes inter-individual differences in cognitive preconditions of the learning group.	4	4
(2) Description of the learning task	The learning task explicitly comprises different cognitive levels (explicit instruction of student differentiation).	4	3
(3) Applying descriptions to the specific situation	The teacher describes the specific cognitive levels of students (student differentiation) towards the learning tasks following the „zone of proximal development“.	2	2
(4) Connecting elements of planning	Learning task(s) and lesson objective(s) is/ are connected.	3	1

Figure 4 Coding scheme for analyzing pedagogical adaptivity in written plans of demonstration lessons using generic and subject-specific indicators

4 Summary of Project Findings

4.1 Generic and Subject-Specific Lesson Planning

Figure 5 shows the distribution of item threshold parameter from one-dimensional IRT scaling, indicated by a circle for each item. To facilitate reading, the distribution of items is split up into generic item thresholds (left side) and into subject-specific item thresholds (right side). Each specific median is indicated by a rectangle. Subject-specific code thresholds are generally in the upper range of item difficulty. The median of their threshold values is about one logit higher ($Mdn = 1.03$) than the median of the threshold values of generic codes ($Mdn = -.13$). Using the *Mann-Whitney U test* as a non-parametric test, mean difference is significant (asymptotic p [two-tailed] = .042).

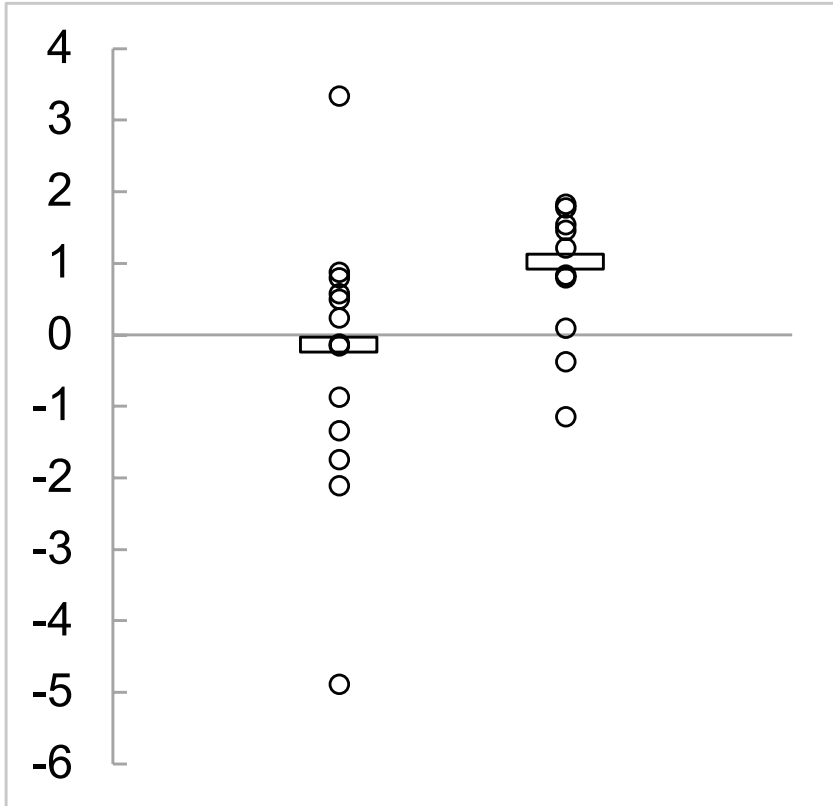


Figure 5 Item threshold parameter (circles) and median (rectangle) from one-dimensional IRT-scaling, split into generic indicators (left side) and subject-specific indicators (right side)

4.2 Learning Progress during Induction

We used ability estimates from one-dimensional IRT scaling to indicate generic and subject-specific planning skills demonstrated in the written plans to time point one (first plan) and time point two (last plan). We used the item threshold parameter medians to cut the ability continuum into three sections: High ability estimates reflect both generic and subject-specific lesson planning skills (Level II), moderate ability estimates reflect generic skills only (Level I), and low estimates

do not sufficiently fulfill the requirements defined by our coding scheme (Figure 4). As Figure 6 illustrates, at the beginning of induction, less than 10 % of lesson plans fulfill subject-specific requirements and about two third of all plans do not even reach the generic requirements sufficiently. At the end of induction, about 40 % of lesson plans reach the level of subject-specific lesson planning demands and there is hardly any plan not sufficiently fulfilling generic requirements (less than 5 % below Level I).

This gain in planning skills over a time span of about 1.5 years can be confirmed using continuous scores (for more details, see König et al. 2019). Scores for the last written plan are almost two standard deviation higher than for the first plan. Using the panel sample of pre-service teachers who not only had submitted their lesson plans from two time points, but who could be matched ($n = 116$), mean differences are significant ($t[1,115] = 13.31, p < .001$) and practically relevant ($d = 1.6$). These statistical findings altogether show a substantial learning progress in lesson planning skills among pre-service language teachers during the second phase of initial teacher education (induction phase).

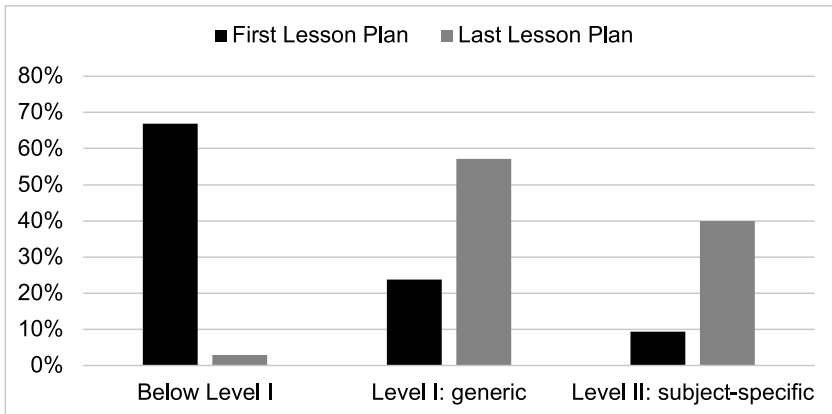


Figure 6 Distribution on planning competence level at each time point

4.3 Factors Influencing the Planning Skills

To analyze potential factors influencing planning skills, two linear regression analyses were conducted, one for each time point (for details, see König et al. 2019). At time point 1, 41 % variance of planning competence as the dependent variable could be explained. At time point 2, 19 % variance of planning competence could

be explained. Regarding learning opportunities during induction, a scale measuring planning-aspects requirements turned out to influence planning skills significantly at both time point 1 ($\beta = .37, p < .001$) and time point 2 ($\beta = .18, p < .05$). The scale refers to aspects that the pre-service teachers' teacher educators required them to cover in their written plans. It comprises five items (e.g., "accounting for learning dispositions for students" and "providing instruction allowing inner differentiation of students"), each with four categories of agreement (using a Likert scale). The more concrete the requirements are – according to which, a pre-service teacher had to write his or her lesson plan – the better the skills he or she provides in the written plan. The plan length also significantly predicts planning skills when looking at the first demonstration lesson plan ($\beta = .23, p < .01$). We could not find other important variables such as teaching experience or teaching type (for more details on teaching types in our sample, see Section 3.3) as being significant predictors for planning skills. On the level of professional knowledge, the pre-service teachers' pedagogical knowledge significantly predicts planning skills at the first time point ($\beta = .19, p < .01$).

5 Discussion and Outlook

Lesson planning of teachers as a research field has received little attention so far in terms of modelling and measuring relevant competences. However, as lesson planning constitutes a substantial part of a teacher's daily work and teacher education provides relevant learning opportunities for future teachers to develop correspondent planning skills, teacher competence research in this area is clearly needed. In the PlanvoLL-D project, we developed and applied a standardized method for analyzing written lesson plans, which highlights the demand for pedagogical adaptivity – both on a generic and on a subject-specific level. We investigated a competence model and the measurement of this planning competence skill using a database of more than 300 written plans of pre-service teachers' demonstration lessons from two time points during induction. Out of this material, we reconstructed planning decisions and created indicators that served to quantify teacher candidates' skill of adaptive lesson planning. Findings show that pre-service language teachers are more highly challenged with subject-specific lesson planning than with generic lesson planning. During induction, pre-service teachers' skill of adaptive lesson planning increases significantly and to a large extent (large effect). Certain factors influencing pre-service planning skill such as pedagogical knowledge acquired during university or learning opportunities during induction could be identified. These findings may enrich the discussion on learning adaptive teaching (König et

al. 2019), and they also inform about the validity of the measurement instrument capturing situation-specific lesson planning skills.

One major limitation of our approach might be that the written plans are part of the examination and certification process during the German induction phase. To some extent it still remains an open question how pre-service teachers' planning decisions are influenced by institutional requirements or teacher educators' preferences regarding demonstration lessons. Moreover, planning is not only restricted to a single lesson. Even pre-service teachers during induction in Germany are required to plan units of lessons. Our measurement approach does not fully account for such a long-term scope of planning.

The research agenda of the PlanvoLL-D project started with a particular focus on pedagogical adaptivity, i.e., the ways in which the assignments of the respective lesson fit with the cognitive level of the learning group (König et al. 2015; 2017b; König 2019). Taking this as a central demand that teachers have to master, our competence model comprises both generic and subject-specific aspects. Adaptivity is, however, only one aspect of lesson planning. We therefore increased the scope and extent of our lesson planning competence framework by adding another demand: structuring of the lesson, i.e., how a teacher plans the lesson sequencing to fulfill didactic functions and effective time management. We have started analyzing written lesson plans and created indicators to quantify teachers' planning decisions in this new area (e.g., Krepf and König 2019).

As part of the transfer activities of the PlanvoLL-D project, demands of pedagogical adaptivity and structuring the lesson have further been reflected in the development of a test design framework that can be used for a standardized test measuring lesson planning. Currently, such a test development has started as part of the Cologne project funded by the BMBF program for increasing the quality of teacher education (Qualitätsoffensive Lehrerbildung, Project *ZuS – Zukunftstrategie Lehrer*innenbildung Köln*). The test comprises several vignettes, each providing a planning situation as a complex stimulus followed by several test items measuring perception, interpretation, and decision-making in such simulated planning situations. First findings will be available in the near future, therefore continuing the research agenda put forward by the PlanvoLL-D project.

Funding

This work was supported by the Federal Ministry of Education and Research, Germany [Bundesministerium für Bildung und Forschung, BMBF], grant number 01PK15014A, 01PK15014B, 01PK15014C.

References

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Longman.
- Baumert, J. (2016). Leistungen, Leistungsfähigkeit und Leistungsgrenzen der empirischen Bildungsforschung. Das Beispiel von Large-Scale-Assessment-Studien zwischen Wissenschaft und Politik. *Zeitschrift für Erziehungswissenschaft*, 19, Supplement 1, 215–253.
- Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society*, 24(3), 200–212.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Borko, H. & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal*, 26(4), 473–498.
- Bremerich-Vos, A., König, J., & Fladung, I. (2019). Fachliches und fachdidaktisches Wissen von angehenden Deutschlehrkräften im Referendariat: Konzeption und Ergebnisse einer Testung in Berlin und NRW. [Content knowledge and pedagogical content knowledge of trainee German language teachers: design and results of a test applied in Berlin and North Rhine-Westphalia.] *Zeitschrift für Empirische Hochschulforschung*.
- Bromme, R. (1981). *Das Denken von Lehrern bei der Unterrichtsvorbereitung: Eine empirische Untersuchung zu kognitiven Prozessen von Mathematiklehrern*. [Teachers' thoughts for planning: An empirical study on cognitive processes of mathematics teachers.] Weinheim: Beltz.
- Buchholtz, C. & König, J. (2015). Erfassung von Planungskompetenz im Praxissemester. *Journal für LehrerInnenbildung*, 15(1), 39–45.
- Clark, C. M., & Peterson, P. L. (1986). Teachers' thought processes. In M. Wittrock (Eds.), *Handbook of research on teaching* (pp. 255–296). New York, NY: Macmillan.
- Cochran-Smith, M., & Villegas, A. M. (2016). Research on teacher preparation: Charting the landscape of a sprawling field. In D. H. Gitomer & C. A. Bells (Eds.), *Handbook of Research on Teaching* (5th edition, pp. 439–547). Washington, DC: AERA.
- Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review*, 18(3), 237–278.
- Corno, L. Y. N. (2008). On teaching adaptively. *Educational Psychologist*, 43(3), 161–173.
- Darling-Hammond, L., Newton, S. P., & Wei, R. C. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability*, 25(3), 179–204.
- Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Eds.), *The road to excellence: The acquisition of expert performance in the arts and science, sports, and games* (pp. 1–50). Mahwah, NJ: Lawrence Erlbaum.
- European Commission (2013). *Supporting teacher competence development for better learning outcomes*. Brussels, Belgium: European Commission.

- Glutsch, N., Bremerich-Vos, A., Buchholtz, C., König, J., Fladung, I., Lammerding, S., Strauß, S., & Schleiffer, C. (2019). PlanvoLL-D – Die Bedeutung des professionellen Wissens angehender Deutschlehrkräfte für ihre Planung von Unterricht: Validierung und methodische Innovation. Skalendarisierung zu Instrumenten der Ausbildungsinhalte und Schulpraxis, Messzeitpunkte 1 und 2, Sommer 2016 und Winter 2017/18. Dokumentation. Köln: Universität zu Köln.
- Hardwig, T., & Mußmann, F. (2018). Zeiterfassungsstudien zur Arbeitszeit von Lehrkräften in Deutschland. Konzepte, Methoden und Ergebnisse von Studien zu Arbeitszeiten und Arbeitsverteilung im historischen Vergleich. Expertise im Auftrag der Max-Träger-Stiftung. Göttingen.
- Hill, J., Yinger, R., & Robins, D. (1983). Instructional planning in a laboratory preschool. *The Elementary School Journal*, 83(3), 182–193.
- Housner, L. D., & Griffey, D. C. (1985). Teacher cognition: Differences in planning and interactive decision-making between experienced and in-experienced teachers. *Research Quarterly for Exercise and Sport*, 56(1), 45–53.
- Jacobs, C. L., Martin, S. N. & Otieno, T. C. (2008). A science lesson plan analysis instrument for formative and summative program evaluation of a teacher education program. *Science Education*, 92(6), 1096–1126.
- John, P. D. (2006). Lesson planning and the student teacher: re-thinking the dominant model. *Journal of Curriculum Studies*, 38(4), 483–498.
- Johnson, K. (2003). *Designing language tasks*. Hampshire, New York: Palgrave Macmillan.
- Jong, T. d., & Ferguson-Hessler, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist*, 3(2), 105–113.
- Kaiser, G., Busse, A., Hoth, J., König, J., & Blömeke, S. (2015). About the Complexities of Video-Based Assessments: Theoretical and Methodological Approaches to Overcoming Shortcomings of Research on Teachers' Competence. *International Journal of Science and Mathematics Education*, 13(2), 369–387.
- Kaiser, G., & König, J. (2019). Competence Measurement in (Mathematics) Teacher Education and Beyond: Implications for Policy. *Higher Education Policy*, 32, 1–19.
- Kang, H. (2017). Preservice teachers' learning to plan intellectually challenging tasks. *Journal of Teacher Education*, 68(1), 55–68.
- Kärner, T., Bonnes, C., & Schölzel, C. (2019). Bewertungstransparenz im Referendariat [Assessment Transparency in Teacher Training]. *Zeitschrift für Pädagogik*, 65(3), 378–400.
- König, J. (2019). PlanvoLL-D: Planungskompetenz von angehenden Lehrerinnen und Lehrern im Fach Deutsch. In N. McElvany, W. Bos, H. G. Holtappels, & A. Ohle-Peters (eds.), *Bedingungen und Effekte von Lehrerbildung, Lehrkraftkompetenzen und Lehrkrafthandeln* (S. 67–85). Münster: Waxmann.
- König, J. & Blömeke, S. (2013). Preparing Teachers of Mathematics in Germany. In J. Schwille, L. Ingvarson & R. Holdgreve-Resendez (eds.), *TEDS-M Encyclopaedia. A Guide to Teacher Education Context, Structure and Quality Assurance in 17 Countries. Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)* (pp. 100–115). Amsterdam: IEA.
- König, J., Blömeke, S., Paine, L., Schmidt, B. & Hsieh, F.-J. (2011). General Pedagogical Knowledge of Future Middle School Teachers. On the Complex Ecology of Teacher Education in the United States, Germany, and Taiwan. *Journal of Teacher Education*, 62 (2), 188–201.

- König, J., & Bremerich-Vos, A. (2019). Deutschdidaktisches Wissen angehender Sekundarstufenlehrkräfte: Testkonstruktion und Validierung. [German pedagogical content knowledge of future secondary teachers: test construction and validation]. *Diagnostica*.
- König et al. (2017a) = König, J., Bremerich-Vos, A., Buchholtz, C., Lammerding, S., Strauß, S., Fladung, I. & Schleiffer, C. (2017). Modelling and validating the learning opportunities of preservice language teachers: On the key components of the curriculum for teacher education. *European Journal of Teacher Education*, 40(3), 394–412.
- König et al. (2017b) = König, J., Bremerich-Vos, A., Buchholtz, C., Lammerding, S., Strauß, S., Fladung, I. & Schleiffer, C. (2017). Die Bedeutung des Professionswissens von Referendarinnen und Referendaren mit Fach Deutsch für ihre Planungskompetenz (PlanvoLL-D). In S. Wernke & K. Zierer (Hrsg.), *Die Unterrichtsplanung: Ein in Vergessenheit geratener Kompetenzbereich?! Status Quo und Perspektiven aus Sicht der empirischen Forschung* (S. 121–133). Bad Heilbrunn: Klinkhardt.
- König, J., Bremerich-Vos, A., Buchholtz, C., Fladung, I., & Glutsch, G. (2019). Pre-service teachers' generic and subject-specific lesson-planning skills: On learning adaptive teaching during initial teacher education. *European Journal of Teacher Education*. doi: 10.1080/02619768.2019.1679115.
- König, J., Buchholtz, C. & Dohmen, D. (2015). Analyse von schriftlichen Unterrichtsplanungen: Empirische Befunde zur didaktischen Adaptivität als Aspekt der Planungskompetenz angehender Lehrkräfte. *Zeitschrift für Erziehungswissenschaft*, 18(2), 375–404.
- Krepf, M., & König, J. (2019). Strukturierung bei der Unterrichtsplanung als Voraussetzung für Klassenführung im Unterricht. [Structuring when lesson planning as a requirement for classroom management.] Presentation on the conference of Gesellschaft für Empirische Bildungsforschung (GEBF), Cologne, Germany, 25–27 February 2019.
- Leinhardt, G., & Greeno, J. (1986). The cognitive skill of teaching. *Journal of Educational Psychology*, 78(2), 75–95.
- Masters, G. N. (2016). Partial credit model. In van der Linden, W. (ed.), *Handbook of Item Response Theory, Volume One: Models* (ch. 7, pp. 137–154). Chapman and Hall/CRC.
- Mutton, T., Hagger, H., & Burn, K. (2011). Learning to plan, planning to learn: The developing expertise of beginning teachers. *Teachers and Teaching*, 17(4), 399–416.
- Neubrand, M., Jordan, A., Krauss, S., Blum, W., & Löwen, K. (2013). Task analysis in CO-ACTIV: Examining the potential for cognitive activation in German mathematics classrooms. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 125–144). Boston, MA: Springer.
- Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., Pierczynski, M., & Allen, M. (2018). Teachers' Instructional Adaptations: A Research Synthesis. *Review of Educational Research*, 88(2), 205–242.
- Pecheone, R., & Chung, R. (2007). Technical report of the Performance Assessment for California Teachers (PACT): Summary of validity and reliability studies for the 2003–04 pilot year. Stanford: Stanford University.
- Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal*, 24(1), 13–48.
- Scholl, D. (2018). *Metatheorie der Allgemeinen Didaktik: Ein systemtheoretisch begründeter Vorschlag*. [Metatheory of general didactics: A systemtheoretically founded approach.] Bad Heilbrunn: Julius Klinkhardt.

- Schoenfeld, A. H. (1998). Toward a theory of teaching-in-context. *Issues in Education*, 4(1), 1–94.
- Shavelson, R. J., & Borko, H. (1979). Research on teachers' decisions in planning instruction. *Educational Horizons*, 57(4), 183–189.
- Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of educational research*, 51(4), 455–498.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Research*, 57(1), 1–22.
- Smith, T. W., & Strahan, D. (2001). Toward a prototype of expertise in teaching. A descriptive case study. *Journal of Teacher Education*, 55(4), 357–371.
- Stigler, J. W., & Miller, K. F. (2018). Expertise and Expert Performance in Teaching. In A. Ericsson, R.R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (2nd edition, Ch. 24, pp. 431–452). Cambridge University Press.
- Strietholt, R., & Terhart, E. (2009). Referendare beurteilen. Eine explorative Analyse von Beurteilungsinstrumenten in der zweiten Phase der Lehrerbildung. [Assessing pre-service teachers. An explorative analysis of assessment instruments during the second phase of teacher education.] *Zeitschrift für Pädagogik*, 55(4), 622–645.
- Taylor, P. H. (1970). *How teachers plan their courses*. Slough, Berkshire: National Foundation for Educational Research.
- Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge: Cambridge University Press.
- Westermann, D. A. (1991). Expert and novice teacher decision making. *Journal of Teacher Education*, 42(4), 292–305.
- Yinger, R. J. (1977). *A study of teacher planning: Description and theory development using ethnographic and information processing methods* (Unpublished doctoral thesis), Michigan State University, USA.
- Zahorik, J. A. (1975). Teachers' planning models. *Educational Leadership*, 33(2), 134–139.



2.4

Relationships between Domain-Specific Knowledge, Generic Attributes, and Instructional Skills

Results from a Comparative Study with Pre- and In-Service Teachers of Mathematics and Economics

Kuhn, C., Zlatkin-Troitschanskaia, O., Lindmeier, A., Jeschke, C., Saas, H., and Heinze, A.

Abstract

We introduce a theoretical framework on teachers' instructional skills to describe how they can be modeled across different domains. This framework conceptualizes teachers' instructional skills as action-related skills (during instruction) and reflective skills (before and after instruction), which are considered crucial for coping with the practical demands of everyday teaching in a specific subject. The theoretical framework assumes that both skill facets are influenced by the teacher's professional knowledge, generic attributes such as general cognitive abilities or ambiguity tolerance as well as affective and motivational factors. To investigate the relationships between teachers' instructional skills, domain-specific knowledge and generic attributes across different domains, the analytical model focuses on two subjects, mathematics and economics. Based on our study with pre- and in-service teachers of mathematics and economics ($N = 564$), which for the first time considers two subjects, we present results on these relationships. The findings are discussed with regard to their transferability to other domains.

Keywords

Pre- and in-service teachers, instructional skills, mathematics, economics, generic attributes, pedagogical content knowledge, content knowledge

Acknowledgements and Funding Information

The study was funded by the German Federal Ministry of Education and Research with the funding numbers 01PK15012. We would like to thank the reviewers for their constructive and valuable feedback.

1 Introduction

The daily routine of teaching requires teachers to meet a large variety of demands. The requirements in- and outside of the classroom are characterized by a high degree of situativity, contextuality, multidimensionality, simultaneity and immediacy, which make them particularly complex (Baxter and Lederman 1999; Oser et al. 2009; Borko and Shavelson 1990; Jackson 1990). Given the complexity of demands, the necessary knowledge and skills among the students of teacher education should be promoted as early as during their university studies to build up the necessary foundations for their profession at an early stage (Darling-Hammond and Lieberman 2012).

The prerequisites for teachers to meet these requirements have been intensively discussed in research. The early discourse of expertise research demonstrates that the knowledge base of a teacher consists of more than propositional knowledge. Rather, further forms of knowledge representations are necessary when describing and explaining a teacher's performance (Carter 1990; Darling-Hammond et al. 2013; Fenstermacher 1994; Schön 1983). Shulman (1986b) assumes that in addition to propositional knowledge, case and strategic knowledge are necessary to flexibly meet the demands of teaching practice. In competence research and according to the established definition of competences by Weinert (2001), the approaches to explaining teacher prerequisites are expanded to include general personal attributes, motivation and self-efficacy (e.g., Baumert and Kunter 2006). Based on a revision of the previous approaches, recent research emphasizes the importance of specific skills to explain the connection between teachers' dispositions and their real actions in instructional practice (Blömeke et al. 2015a; Kersting et al. 2012; Seidel and Stürmer 2014; Zlatkin-Troitschanskaia et al. 2019a).

Based on the various theoretical approaches and models to describe the necessary teacher prerequisites, analyses have been carried out over the years to empirically map parts of their relationships (Section 3). An empirical examination shows how challenging it is to map an overall picture in which the interdependent correlations of the individual components are considered comprehensively (Sadler 2013; Shulman 1986a). Increasingly complex theoretical models place high demands on the measurement methodology (Zlatkin-Troitschanskaia and Pant 2016). It is not surprising that there is still no uniform, empirically tested overall picture of the relationship of teachers' prerequisites in teacher research.

This also results in the still unanswered question of the domain-specificity of teachers' competences. Although the distinction of knowledge according to Shulman (1986b) in general pedagogical knowledge (GPK), content knowledge (CK), and pedagogical content knowledge (PCK) is widespread in research and has also been empirically confirmed for various subjects (Depaepe et al. 2013; Kuhn et al. 2014; Riese and Reinhold 2012), their relationship to skills that are more closely related to action has so far only been rudimentarily researched (Blömeke et al. 2016; Kersting et al. 2012). With regard to the skills themselves, there are different assumptions as to whether they can be regarded as domain-specific (e.g., "usable knowledge" in mathematics, Kersting et al. 2012) or generic ("professional vision", Seidel and Stürmer 2014; "perception, interpretation, decision-making", Blömeke et al. 2015a; Santagata and Yeh 2016). Furthermore, the question remains to what extent generic, non-subject-specific attributes of teachers (e.g., general cognitive abilities) account for correlations between subject-specific teacher skills. The patterns of influence might vary between subjects, as the nature of the subjects and their disciplinary structure differs (e.g., Gess-Newsome 1999; Shulman 1986a). The majority of studies to date still focus on just one subject respectively; studies involving several subjects have been the exception so far (for teachers from Germany, Niermann 2017). To examine the relationships across subjects, a comparative approach involving various subjects under control of generic factors is needed.

In Germany, teacher education has three stages: The first, theoretical stage takes place at university and comprises three years of bachelor studies and two years of master studies. The second stage consists of one and a half to two years of supervised practical training at schools. The third stage involves professional, fully autonomous teaching at schools. In general, teachers' instructional skills should increase during all three stages, university education, practical training, and professional practice (Berliner 1995, see also Buschang et al. 2012). However, discussion on the conditions of 'deliberate practice' suggests that simply 'doing a job' does not necessarily lead to a higher level of expertise (Ericsson 2000; Ericsson et al. 1993). Instead, the level of expertise increases particularly with opportunities

for structured learning, for example through guided self-reflection and feedback (e.g., Bronkhorst et al. 2014). Teachers encounter such opportunities especially during the first and second stage of teacher education (for empirical results, e.g., Kleickmann et al. 2013; Hill et al. 2005; Nilsson and Loughran 2012), since these two stages include both a practical aspect and university courses specifically on subject-related didactics. In the third stage (autonomous teaching), such structured learning opportunities can only be found in the form of subject-related didactic advanced vocational training, which is rather rare.

In our paper, we examine the relationships between instructional skills, domain-specific knowledge and generic attributes of teachers of mathematics and economics. In addition, we focus on the level of teachers' instructional skills at all three stages of teacher education (teacher students, trainee teachers and in-service teachers) for both subjects.

Based on the state of research, we introduce a framework on teachers' instructional skills to theoretically depict the connection between dispositions, skills and performance. The theoretical framework conceptualizes teachers' instructional skills as action-related skills (AS) and reflective skills (RS), which are considered crucial during instruction (AS) as well as before and after instruction (RS) in a specific subject. The framework assumes that both skill facets are influenced by the teacher's professional knowledge, generic attributes such as general cognitive abilities, ambiguity tolerance as well as affective and motivational factors, along with socio-biographical characteristics. Since our framework is based on previous findings from teacher research, we present selected empirical findings on the relationships between the facets of instructional skills to develop our research hypotheses.

To investigate the relationships and the levels of instructional skills across domains, our comparative study includes pre- and in-service teachers of two subjects, mathematics and economics ($N = 564$). Both subjects can be considered well-structured teaching domains (Short 1995). Mathematics and economics are two different but related subjects as mathematics is applied to solve certain problems in economics (CEE 2010). Studies also show that individuals' performance in mathematics tests and their performance in economics tests correlate (e.g., Ballard and Johnson 2004; Shavelson et al. 2019a). This is not surprising, as mathematics can be conceptualized as one facet of economic knowledge and skills, and therefore the two domains can be considered related disciplines (Shavelson et al. 2019b).

For the first time, this study compares two subjects and enables first evidence-based insights into the relationships between generic and domain-specific facets of teachers' instructional skills. By considering three status groups (stu-

dents, trainee teachers, experienced teachers), we can infer first conclusions about differences in the groups' respective level of instructional skills. On the basis of these findings, important insights can be gained for fostering these skills in teacher education and training programs. The results will also be discussed with regard to their transferability to other domains.

2 Theoretical Framework on Teachers' Instructional Skills

The existing modeling approaches for the description of teacher prerequisites can be divided into analytical and holistic approaches. Analytical approaches focus on single facets required for professional performance, i.e. professional competences are influenced by different interrelated cognitive and non-cognitive abilities and traits that can be measured separately and are necessary for competent behavior in professional contexts (Shulman 1986b; Schön 1983). The holistic modeling approach considers all cognitive and non-cognitive resources and their interaction, and competences are understood as a complex superordinate aptitude that enables a teacher to master specific professional demands (Corno and Snow 1986, Shavelson et al. 2019b; Zlatkin-Troitschanskaia et al. 2019b).

Analytical modeling approaches deal with several challenges such as evermore complex models which can hardly be operationalized empirically at this level of detail. As a consequence, analytical modeling approaches often focus on particular cognitive characteristics of teachers, for instance, on de-contextualized declarative knowledge (e.g., Kunter et al. 2011). Only parts of single facets of teachers' knowledge have been empirically assessed so far; this typically includes CK and PCK using traditional test methods (e.g., multiple-choice tests, text vignettes, or teacher reflections) (e.g., Holtsch et al. 2018). Yet, even these facets are only to a limited extent suitable for describing situated professional action of teachers in the instructional situation. To counterbalance the shortcomings of existing approaches, the current next generation of performance-oriented assessments based on a holistic view might be used to complement analytical approaches (Darling-Hammond et al. 2013; Jeschke et al. 2019b; Kuhn et al. 2018; Shavelson et al. 2019b; Zlatkin-Troitschanskaia et al. 2019b).

One holistic approach stems from the American tradition of performance assessment in education (Shavelson et al. 2019b), which emerged in research on "adaptive action" (Corno and Snow 1986). Since skills in real life are not neatly divided into single components, they can be analyzed through the stages that the individual goes through while handling a challenge (Shavelson et al. 2019b; Zlatkin-Troitschanskaia et al. 2019b). Current holistic approaches for compe-

tence modeling focus on competences close to professional actions, and encompass all abilities, skills and attitudes that are important for mastering professional demands (Weinert 2001). Performance assessments are based on a criterion sampling measurement approach that focuses on sampling real-life events and consolidating them into frameworks (Shavelson et al. 2019b; Zlatkin-Troitschanskaia et al. 2019b). One implication for assessment is that competence modeling must be based on a detailed analysis of real professional teaching requirements considering the complexity and contextual nature of real classroom instruction (Darling-Hammond and Baratz-Snowden, 2005; Oser et al. 2009). Blömeke et al. (2015a) differentiate several dispositions (i.e., professional knowledge, affect/motivation and generic attributes) as the basis for situation-specific skills (i.e., perception, interpretation, decision-making) and teaching performance (Zlatkin-Troitschanskaia et al. 2019a).

Based on these conceptual considerations and focussing on the holistic modeling approach, we developed our theoretical framework on teachers' instructional skills (Figure 1; Lindmeier 2011; Zlatkin-Troitschanskaia et al. 2019a). We assume that teachers' situation-specific skills (described as perception, interpretation and decision-making by Blömeke et al. 2015a), can differ depending on two essential facets of teaching practice, which exist in all subjects: action-related skills (AS) for in-classroom teaching practice, and reflective skills (RS) for pre- and post-classroom demands (Lindmeier 2011; Zlatkin-Troitschanskaia et al. 2019a):

- AS are considered a domain-specific cognitive resource that enables teachers to handle specific subject-related situations during instruction in the classroom, for example, when reacting immediately to students in a fast and adaptive fashion (e.g., a teacher should be able to recognize students' difficulties and to react flexibly in a didactically appropriate manner).
- RS are considered a domain-specific cognitive resource that enables teachers to prepare and evaluate specific situations in pre- and post-instructional phases (e.g., a teacher should already consider how to effectively prevent misconceptions among students while planning the lesson).

In accordance with our assumptions, we divide the dispositions into professional knowledge, various generic attributes as well as affective and motivational factors. The professional knowledge base as well as the affective and motivational factors can be further viewed as both domain-specific constructs (e.g., CK, PCK, motivation to teach a certain subject) and as cross-domain constructs (e.g., general

motivation to teach).¹ Among the generic attributes, we focused on the following dispositions, which can be expected to have a significant relationship with situation-specific skills (for the current state of research, see Section 3): ambiguity tolerance, Big Five personality traits, general cognitive abilities, and teacher-specific self-efficacy. In the following, we present selected empirical findings on the relationships between the various components. The findings serve as a basis for the derivation of our research hypotheses.

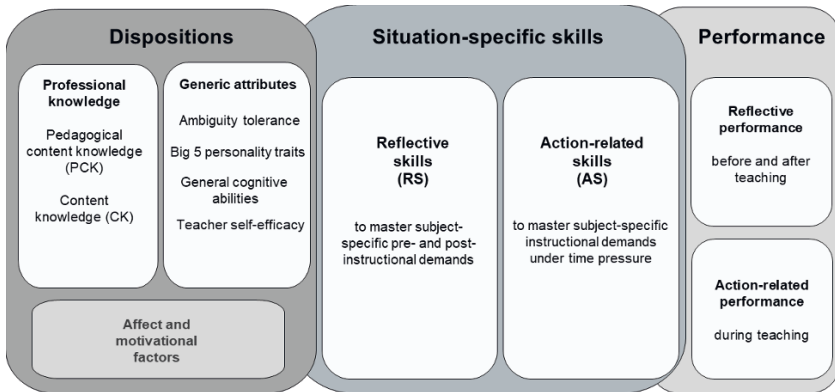


Figure 1 Theoretical framework on teachers' instructional skills (according to Lindmeier 2011; Zlatkin-Troitschanskaia et al. 2019a, p. 154)

3 State of Research

3.1 Professional Knowledge and Situation-Specific Instructional Skills

Research on teachers' knowledge and skills is gaining importance across disciplines and internationally (e.g., Richmond et al. 2019; Gitomer and Bell 2016). As early research shows, the subject-specific knowledge of teachers predicts both instructional quality and student learning (Shulman 1986a; e.g., for the mathe-

¹ Based on the state of research, in our theoretical framework we also differentiate motivational aspects as an important component of teachers' dispositions (Figure 1). Due to the limited test time, however, we were only able to focus on selected aspects in our study, which means these facets were only assessed with a few questions in the socio-demographic part, and we will therefore not go into them in more detail in this article.

mathematical knowledge of mathematics teachers, Hill et al. 2005; Lindmeier et al. under review). Recent studies have indicated that knowledge alone is not sufficient to explain teachers' performance and to describe the situational prerequisites of teachers to cope with typical teaching demands (Kersting et al. 2012; Santagata and Sandholtz 2018). Studies in the domain of mathematics confirmed that teachers' instructional skills are empirically separable from knowledge (Blömeke et al. 2014; Knievel et al. 2015; Hepberger et al. 2019) and even more predictive for instructional quality and student learning than teacher knowledge (Kersting et al. 2010; Kersting et al. 2012). Blömeke et al. (2014) also showed that mathematics teachers' CK, acquired during teacher education, was a crucial predictor for performance characteristics such as their perception of classroom situations or how quickly they recognize student difficulties. As König et al. (2014) demonstrate in their study, *general* pedagogical knowledge and skills can be empirically distinguished as well.

3.2 Generic Attributes and Situation-Specific Instructional Skills

Studies are concerned with identifying generic teachers' attributes that favor or inhibit professional teaching skills (Bromme and Haag 2008; Keller-Schneider 2009; Lin et al. 2005; Lohmann et al. 1966). In the following, we focus on teachers' ambiguity tolerance, the Big Five inventory, general cognitive abilities and teachers' self-efficacy.

Ambiguity tolerance. One focal point of many studies is the role of ambiguity tolerance as a personal trait that determines differences in dealing with uncertainty (e.g., Sorrentino et al. 1984). Ambiguity-tolerant persons tolerate uncertain situations and even have a real need for them, as they interpret such situations as challenges (König 2003). Many of the demanding situations of everyday teaching practice can be characterized as particularly uncertain and complex, for instance, student learning processes or the challenge of engaging with a new class (Dalbert and Radant 2010). Due to this openness of teaching routine, teachers generally (inter)act in uncertain situations. The low predictability of teaching is seen by ambiguity-tolerant individuals as an opportunity to give students space for independent constructions of the subject matter: The more pronounced a teacher's tolerance for ambiguity is, the more open he or she in turn designs his or her teaching practice (König and Dalbert 2007). A study with teachers at vocational schools confirmed that a more positive perception of one's own performance as well as more frequent use of cooperative learning methods are common characteristics of

ambiguity-tolerant teachers, and the general ability to adequately fulfil pedagogical requirements increases with higher ambiguity tolerance (König 2003). Mayr (2011) also identified ambiguity tolerance as a “special” personal characteristic that is positively related to pedagogical skills.

Big Five personality traits. The most influential model on the construct of personality traits is the Big Five personality model, which differentiates between five traits to describe differences in behavior, thoughts, motivations, and emotions: openness, conscientiousness, extraversion, agreeableness, and neuroticism (John et al. 2008). Studies on personality traits have so far lacked a consistent picture of teacher personality (Eulenberger 2015). The use of the Big Five inventory (Benet-Martínez and John 1998) has led to heterogeneous findings of the relationship between a teacher’s personality and teaching behavior (Bastian et al. 2017; Cutchin 1998; Job 2004; Rockoff et al. 2011). Klassen and Tze (2014) reviewed a meta-analysis (43 studies; $n = 9,216$ teachers) in which all Big Five personality traits, except for agreeableness, were shown to significantly correlate with teacher effectiveness. In contrast, Corcoran and O’Flaherty (2018), who used performance rankings resulting from classroom observations of 400 pre-service teachers to assess teaching performance, found no significant relationship between Big Five personality traits and teaching performance, whereas previous teaching performance in combination with academic achievement scores emerged as significant predictors of teaching performance. Aydin et al. (2013) reported not only significant positive effects on teaching competences for conscientiousness, extraversion, and agreeableness, but also significant negative effects for neuroticism for 206 pre-service teachers. Mayr (2016) revealed extraversion and openness as being particularly relevant to teachers.

General cognitive abilities. Empirical research has revealed that a person’s general cognitive abilities are decisive for academic and professional success and facilitate the acquisition of professional knowledge and skills (Kuncel et al. 2004; Colquitt et al. 2000). General cognitive skills are also decisive for the professional success of teachers, in particular when beginning a career (Kennedy et al. 2008). Some studies report positive correlations between students’ performance (Aloe and Becker 2009; Zumwalt and Craig 2005) and teachers’ diagnostic skills (Kaiser et al. 2012). In two studies with different settings, Kaiser et al. (2012) investigated the relationship between prospective teachers’ cognitive abilities and their accuracy of judgement in the grading of student performance. The Advanced Progressive Matrices Test (APM) by Raven (1962) was used to measure the teachers’ cognitive abilities. In both studies, expectations were confirmed, and cognitive abilities correlated positively with accuracy in student performance evaluation (Kaiser et al. 2012). Furthermore, Mathesius et al. (2019) showed that the three sub-facets of

the IST-2000-R intelligence test (verbal, numerical and figural intelligence) and the sum of the three sub-facets correlate significantly positively with the scientific reasoning of teacher education students in biology.

Teacher' self-efficacy. Self-efficacy is described as a subjective certainty of being able to cope with unknown or difficult situations (Schwarzer and Jerusalem 2002). The self-efficacy of teachers is considered an individual conviction in terms of the extent to which the teacher is able to promote and support the students' learning and behavior, even under difficult conditions (Tschannen-Moran and Hoy 2001). A high expectation of self-efficacy is essentially regarded as positive, since it correlates with a more productive confrontation with challenges, more time for planning lessons, higher motivation, higher stamina, higher use of student feedback for further development of lessons, higher flexibility, more demanding goals, and a high level of achievement (Schwarzer and Warner 2014). Klassen et al. (2011) as well as Tschannen-Moran et al. (1998) provided extensive reviews of teacher efficacy research, showing consistently that teachers' self-efficacy correlates positively with their teaching behavior (Klassen and Tze 2014; see also Ghaith and Yaghi 1997; Guskey 1988; Koşar 2015; Holzberger et al. 2013; Morris-Rothschild and Brassard 2006; Ross 1998; Wolters and Daugherty 2007; Woolfolk et al. 1990). Holzberger et al. (2014) showed a significant correlation between teachers' self-efficacy and three dimensions of instructional behavior (cognitive activation, teacher–student relationship, and classroom management).

3.3 Teachers' Knowledge and Skills in Mathematics and Economics

Most teacher education degrees and programs are mainly designed for a specific subject respectively (e.g., programs for mathematics teachers), although they usually also comprise fundamental general pedagogical topics (e.g., Kunina-Habenicht et al. 2019). Accordingly, most studies of teacher knowledge and skills are limited to one subject, and studies involving several subjects have been the exception so far (for German teachers, Niermann 2017; Praetorius et al. 2015). This is astonishing, as in some educational systems, including Germany, secondary school teachers receive equal training in two subjects, for example, mathematics and physics or mathematics and economics.

Blömeke and colleagues (2016) attempted to empirically separate teacher skills of one domain (mathematics) from general pedagogical skills that are not related to a domain but are operationalized in a similar way (for the context of classroom management). Their results indicated that, in a sample of practicing mathemat-

ics teachers, skills for applying mathematical knowledge are more closely related to skills for applying pedagogical knowledge than to mathematics CK and PCK, giving first evidence that teachers' skills for applying CK and PCK may not be specific to the domain of mathematics. Like most currently available studies, this too focused on teacher skills in only one domain, neglecting to compare teachers' intra-individual ability to apply knowledge in more than one domain even though teachers usually teach two different subjects.

3.4 Development of Teachers' Instructional Skills in Mathematics and Economics

According to teaching expertise research, practical teaching experience is related to a higher development of instructional skills (Baer et al. 2007; Beck et al. 2008). For economics, in particular, Kuhn (2014) shows the expected increase of PCK along the subgroups of bachelor students, master students, trainee teachers, and in-service teachers. The differences in PCK levels are significant, with the exception of the difference between trainee teachers and in-service teachers (for development of E-CK and E-PCK, Seifried and Wuttke 2015). For mathematics, Schönfeld and Kilpatrick (2008) show that M-PCK is increasing upon professional entry due to practical experience, and that the repertoire of teaching strategies is expanding, but M-CK is only being expanded in parts (Llinares and Krainer 2006). The quasi longitudinal comparisons by Kleickmann et al. (2013) show corresponding differences in M-CK and M-PCK between prospective teachers at the end of their training and experienced teachers. Since experienced teachers have an average of 21 years' work experience, it is not possible to make any statements about the changes that may occur during professional practice.

4 Research Framework

4.1 Hypotheses

On the basis of the current state of research and with a focus on the relationships between teachers' instructional skills, domain-specific knowledge and generic attributes as well as on the levels of teachers' instructional skills, we address the following hypotheses in our study for two subjects, economics and mathematics:

- H1: The four constructs (CK, PCK, AS, RS) from the theoretical framework (Figure 1) are related, but empirically separable in each subject.
- H2: Generic attributes (general cognitive abilities, self-efficacy, ambiguity tolerance, neuroticism) show less influence on AS and RS than domain-specific knowledge (CK and PCK) and the patterns of relation are comparable across the two subjects.
- H3: With an increasing degree of domain-specific expertise, a teacher's CK, PCK, AS, and RS become more pronounced.

4.2 Design and Sample

We conducted a comparative, quasi-experimental study with pre- and in-service teachers of two domains, mathematics and economics, including participants who teach both. The combination of the two domains is particularly attractive for teachers at upper-secondary schools with a vocational focus (“Berufsschule”). To achieve quasi-experimental variation and to examine interdependencies between the two domains, the overall sample ($N = 564$) comprises three status groups which differ in their degree of expertise and also in their training in mathematics or economics as well as in both subjects (Table 1). The recruitment of our sample took place at universities, teacher training (“Referendariat”) colleges, and schools from 52 cities in 10 German federal states. The prerequisite for teacher trainees to participate was that they had to be in the second half of their training program so that they can be considered advanced trainees. Participation was voluntary, and a monetary incentive was offered as compensation. The participants also received automated feedback on their results using a feedback tool. Participants’ ages ranged from 18 to 64 years ($M = 30.1$, $SD = 8.41$), and gender distribution was 46.2% female and 53.8% male participants (Table 1).

Table 1 Description of sample

Expertise/ Domain	Mathematics	Both	Economics	Overall
Students	55	54	180	289
Trainee teachers	90	18	49	157
Experienced teachers	24	27	67	118
Overall	169	99	296	564

4.3 Instruments

AS and RS in mathematics and economics were measured using video-based performance assessments (Jeschke et al. 2019b; Kuhn et al. 2018). The performance assessment for mathematics AS (AS-M) comprises 9 items in which participants have to react directly to the students seen in the videos and help them to solve specific mathematical problems. Similarly, the 7 items used to assess AS in economics (AS-E) focus on the teaching of central curricular content in economics. Participants had to respond verbally and under time pressure. Responses were recorded via microphone. Within the 9 items for mathematics RS (RS-M) and the 7 items for economics RS (RS-E) the participants had to reflect on classroom situations seen in the video and provide possible reasons for students' difficulties or alternative actions. Participants provided written responses. A scoring scheme that was developed for each item describes specific criteria for adequate teacher responses based on a theoretical framework and findings from the pre-test studies (Jeschke et al. 2019b; Kuhn et al. 2018; Zlatkin-Troitschanskaia et al. 2019a).

To adequately assess the knowledge components CK and PCK, we used previously tested and validated instruments with closed and open-ended paper-pencil items for mathematics CK (CK-M, 11 items, Dreher et al. 2018) and PCK (PCK-M, 13 items, Loch et al. 2015) as well as for economics CK (CK-E, 14 items, Zlatkin-Troitschanskaia et al. 2014) and PCK (PCK-E, 11 items, Kuhn et al. 2016).

To assess generic attributes, we used the scale of perceived self-efficacy of teachers (Schmitz and Schwarzer 2000) and the scale of figural intelligence for general cognitive abilities (Liepmann et al. 2007); personality traits were assessed using a Big Five inventory (Benet-Martínez and John 1998; Gerlitz and Schupp 2005) as well as an inventory for measuring ambiguity tolerance developed by Reis (1996).

The open-ended (constructed) responses were transcribed and coded by two trained independent raters with interrater agreements of Cohen's κ between acceptable and very good, based on at least 20% of the open responses (randomly selected) for each item (for mathematics: AS-M: $\kappa = .77-.90$ ($M = .84$); RS-M: $\kappa = .80-1.00$, ($M = .92$); CK-M and PCK-M: $\kappa = .70-1.00$ ($M = .89$), for economics: AS-E: $\kappa = .60-.89$ ($M = .76$); RS-E: $\kappa = .5-.78$ ($M = .66$); PCK-E: $\kappa = .60-.89$ ($M = .78$)).

The internal consistency of all test instruments used was acceptable to good (Cronbach's $\alpha = .60-.84$, see Table 2). The relatively lower reliabilities of the domain-specific constructs can still be considered marginally sufficient in view of the scale lengths and the conceptual heterogeneity of the constructs (e.g., Blömeke et al. 2015b; Hill et al. 2004).

Table 2 Description of the test instruments

Instruments	Number of items	Assessment format	Response format	Cronbach's α (<i>N</i>)	Authors
CK-M	11	Paper-pencil	MC ^a /CR ^b	.62 (247)	Dreher et al. 2018
PCK-M	13	Paper-pencil	MC/CR	.60 (393)	Loch et al. 2015
AS-M	9	Video-based	CR/orally	.60 (244)	Jeschke et al. 2019b
RS-M	9	Video-based	CR	.61 (387)	Lindmeier 2011
CK-E	14	Paper-pencil	MC	.60 (393)	Zlatkin-Troitschanskaia et al. 2014
PCK-E	11	Paper-pencil	MC/CR	.61 (387)	Kuhn 2014; Kuhn et al. 2016
AS-E	7	Video-based	CR/orally	.64 (390)	Kuhn et al. 2018
RS-E	7	Video-based	CR	.61 (384)	Kuhn et al. 2018
Teacher Self-efficacy	10	Questionnaire	Rating scale	.66 (513)	Schmitz and Schwarzer 2000
Figural intelligence	20	Paper-pencil	MC	.78 (563)	Liepmann et al. 2007
Ambiguity tolerance	16	Questionnaire	Rating scale	.77 (527)	Reis 1996
BFI-25	25	Questionnaire	Rating scale		Benet-Martínez and John 1998; Gerlitz and Schupp 2005
Extraversion				.84 (539)	
Agreeableness				.66 (518)	
Conscientiousness				.77 (542)	
Neuroticism				.73 (541)	
Openness				.82 (539)	

Notes. ^aMultiple-Choice; ^bConstructed Response

5 Results

H1: The four constructs (CK, PCK, AS, RS) are related, but empirically separable in each domain.

To examine H1, a correlation analysis using SPSS 23 is conducted. The non-standardized sum scores from the tests are used to depict the domain-specific constructs CK, PCK, AS and RS in mathematics and economics. In Tables 3 and 4, the bivariate Pearson correlations between the respective constructs are illustrated.²

Table 3 Pearson correlations between knowledge and skills constructs in the domain of mathematics

	CK-M	PCK-M	AS-M
PCK-M	.39***	–	–
AS-M	.37***	.32***	–
RS-M	.45***	.42***	.49***

Notes. *** $p < .001$

Table 4 Pearson correlations between knowledge and skills constructs in the domain of economics

	CK-E	PCK-E	AS-E
PCK-E	.33***	–	–
AS-E	.30***	.29***	–
RS-E	.21***	.37***	.33***

Notes. *** $p < .001$

The manifest correlations mostly refer to moderate correlations, therefore, the assumption of empirical separability of the constructs in pair-wise correlations (H1) can be regarded as confirmed. High correlations, which point to a greater proximity of the two constructs, can be seen in the domain of mathematics between CK-M and RS-M and between RS-M and AS-M.

The bivariate correlations between the respective constructs are consistently higher in mathematics ($.32 < r < .49$) than in economics ($.21 < r < .37$). This result confirms previous findings on the relationship between CK and PCK, where the

2 We refer to additional analysis using multivariate linear regression models to examine the relationships controlling common variance between CK and PCK (Jeschke et al. 2019a).

correlations are stronger in the domain of mathematics than in other domains, including economics (Section 2). The explanations discussed include the more “substance/content”-focused orientation of mathematics education and a more specialized expertise, which requires a greater synergy between the constructs in the domain of mathematics.

Differences between the two domains can also be seen in the relationships between CK and AS, as well as between CK and RS. While the correlation between CK and AS is weaker (.37) than the correlation between CK and RS (.45) for the domain of mathematics, the opposite is evident for the domain of economics, where the correlation between CK and AS of $r = .30$ is stronger than the correlation between CK and RS (.21) in this sample.

H2: Generic attributes (general cognitive abilities, self-efficacy, ambiguity tolerance, neuroticism) show less influence on the AS and RS than domain-specific knowledge (CK and PCK) and the patterns of relation are comparable across the two subjects.³

As expected, in both subjects, correlations between domain-specific knowledge and AS are higher compared to the weak or non-existent correlations between generic attributes and AS (Table 5). For both subjects, significant weak correlations were found between AS and general cognitive abilities as well as AS and neuroticism (expected negative correlation), while no significant correlations were found between AS and self-efficacy. For AS in economics, in contrast to mathematics, an additional significant, although rather weak, correlation with ambiguity tolerance was identified.

Table 5 Pearson correlations between generic attributes and AS in mathematics and economics

	AS-M	AS-E
Cognitive abilities	.16*	.12*
Teacher self-efficacy	-.03	.03
Ambiguity tolerance	.06	.17**
Neuroticism	-.16*	-.15**

Notes. * $p < .05$; ** $p < .01$

Due to the stronger domain-specific correlation, we used a multiple linear regression model, which initially only included CK and PCK as predictors of AS. In a

³ In the following, only the results for AS are described, due to limited space.

second step, we also included the generic attributes as additional predictors to analyze whether they have a further influence on AS (Tables 6 and 7).

As expected, for both subjects, domain-specific knowledge (PCK, CK) has a highly significant influence on the dependent variable AS, which remains even after the inclusion of generic attributes ($p < .01$). The addition of generic attributes increases the explanatory power of the models, however, in both domains only one significant relationship between the generic attributes and AS was found. In this respect, the patterns between both subjects can be interpreted similarly. For mathematics, there is a significant negative effect on AS for neuroticism ($\beta = .15, p < .05$), whilst for economics there is a significant positive effect of general cognitive skills on AS ($\beta = .10, p < .05$).

Table 6 Multiple linear regressions on the score in AS-M

Variable	Model 1			Model 2		
	Coefficient B	SE (B)	Beta (β)	Coefficient B	SE (B)	Beta (β)
Constant	3.01***	0.71		7.97**	2.87	
PCK	0.21**	0.07	.20**	0.19**	0.07	.19**
CK	0.26***	0.06	.29***	0.25***	0.06	.29***
Self-efficacy				-1.08	0.749	-.09
Ambiguity tolerance				-0.02	0.42	-.00
General cognitive abilities				0.05	0.06	.06
Neuroticism				-0.49*	0.22	-.15*
R ²	.17***			.20***		
Corrected R ²	.17***			.18***		

Notes. $N = 226$ (model 1)/211 (model 2); * $p < .05$, ** $p < .01$, *** $p < .001$

Table 7 Multiple linear regressions on the score in AS-E

Variable	Model 1			Model 2		
	Coefficient B	SE (B)	Beta (β)	Coefficient B	SE (B)	Beta (β)
Constant	2.51***	0.43		1.23	1.68	
PCK	0.19***	0.05	.21***	0.19***	0.05	.21***
CK	0.26***	0.05	.24***	0.26***	0.06	.24***
Self-efficacy				-0.03	0.44	-.00
Ambiguity tolerance				0.36	0.25	.08
General cognitive abilities				0.07*	0.04	.10*
Neuroticism				-0.21	0.12	-.09
R ²	.14***			.18***		
Corrected R ²	.13***			.17***		

Notes. $N = 377$ (model 1)/358 (model 2); * $p < .05$, ** $p < .01$, *** $p < .001$

H3: With an increasing degree of domain-specific expertise, a teacher's CK, PCK, AS, and RS become more pronounced.

For a comparative analysis of teachers' domain-specific knowledge and skills levels, CK, PCK, AS and RS in the three status groups (students, trainee teachers, and experienced teachers) were considered separately. The basic assumption was that the test results of those constructs increase with a higher level of training and increased professional expertise. To investigate this assumption, the mean score values of the different status groups were compared (Figures 2 and 3). The general effects of group affiliation were described through ANOVA and post-hoc analyses.

The explained variance in knowledge and skill levels through group affiliation was fairly high in economics, especially for the constructs PCK-E (partial $\eta^2 = .207$) and RS-E (partial $\eta^2 = .161$), and with moderate effects for the two other constructs (CK-E: partial $\eta^2 = .021$; AS-E: partial $\eta^2 = .055$). In mathematics, a moderate effect was found for AS-M (partial $\eta^2 = .026$), no effects were identified for the three other constructs (CK-M: partial $\eta^2 = .004$, PCK-M: partial $\eta^2 = .004$, RS-M: partial $\eta^2 = .002$).

The results of the post-hoc tests (Games-Howell adjusted) confirm that a person's PCK-E and RS-E are significantly higher with increasing expertise, and indicated significant differences in mean scores between students and trainee teachers (PCK-E: 1.63**, RS-E: 1.35***), students and experienced teachers (PCK-E: 3.34**, RS-E: 2.63***), and trainee teachers and experienced teachers (PCK-E: 1.72**, RS-E: 1.28**). For AS-E, significant differences were found between stu-

dents and trainee teachers (.98*), and students and experienced teachers (1.47***). For CK-E and AS-M, significant differences were identified between students and experienced teachers only (CK-E: .94*; AS-M: 1.60*).

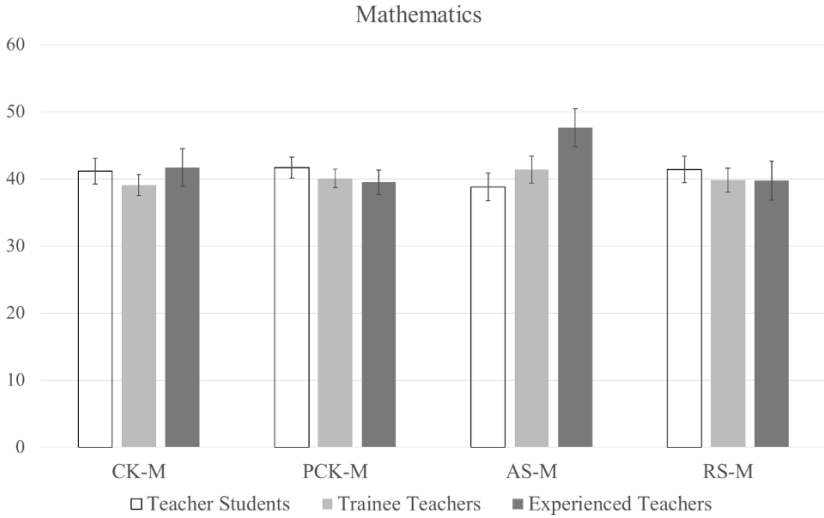


Figure 2 Relative mean scores (CK-M, PCK-M, AS-M, RS-M) by groups in mathematics Error bars represent standard error.

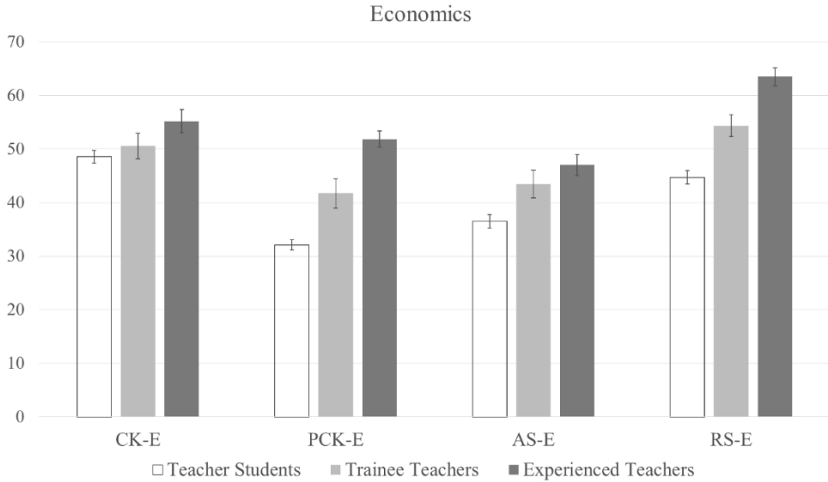


Figure 3 Relative mean scores (CK-E, PCK-E, AS-E, RS-E) by groups in economics. Error bars represent standard error.

6 Discussion and Conclusion

This comparative study investigates how relationships between different facets of instructional skills behave in the domains of mathematics and economics. The results show largely moderate (to strong) correlations between the four constructs (CK, PCK, AS, RS), so that the separability in pairwise correlations (H1) can be confirmed for both domains. In particular, there is a tendency for bivariate correlations to be consistently stronger in mathematics than in economics. The correlations indicate that AS and RS align differently with the conceptual understanding of the subject (for economics, e.g., Mankiw 2012) in mathematics and economics. For economics, spontaneous responses to students' statements in the form of appropriate impulses and explanations (AS) essentially depend on whether the student's mistake was correctly identified by the (prospective) teacher (Zlatkin-Troitschanskaia et al. 2019a), which might be mirrored in these findings. Dealing with students' mistakes reflectively (RS), for instance, when planning future lessons, requires subject-specific understanding as well, but additionally also requires more specific didactical instructions and their application. This may also explain the comparatively strong correlation between PCK-E and RS-E for the domain of economics. For mathematics, the result that subject-specific knowledge

(CK, PCK) tendentially correlates more strongly with RS than with AS supports the assumption that mathematics teachers are able to apply their knowledge more effectively when dealing with reflective tasks such as lesson planning and evaluation (RS) than with instructional tasks under time pressure (AS).

While we identified weak correlations between generic attributes and AS in both domains, it becomes evident that the correlation between domain-specific knowledge (CK, PCK) and AS is not only much stronger, but also partially explains the relationships to generic attributes. In both domains, generic attributes (general cognitive abilities, self-efficacy, ambiguity tolerance, neuroticism) show weaker correlations to AS than the two domain-specific knowledge constructs CK and PCK. Overall, H2 can be regarded as confirmed.

As we assumed in H3, knowledge and skills develop during teacher training, resulting in a higher level of expertise. This hypothesis cannot be confirmed for both domains in our study. Only in economics did participants with a higher level of expertise also achieve higher test scores, thus indicating that the four constructs (CK, PCK, AS, RS) are becoming increasingly more pronounced with the test subjects' level of expertise. In contrast, in mathematics, only an increase in the level of AS could be determined in the given sample.

Since H3 was only tested by cross-sectional (not longitudinal) comparison and not experimental, these findings do not allow any statements about possible causal relationships between the examined constructs. In addition, the instruments still require improvement in terms of reliability, a lack of which may hamper the detection of expertise effects. Despite these limitations, these results provide an important basis for further studies on possible explanatory factors.

With regard to our subsample of pre- and in-service teachers trained in both subjects, mathematics and economics ($n = 99$), we found preliminary evidence suggesting that some of the knowledge and skills used in mathematics are related to the knowledge and skills necessary for teaching economics (Jeschke et al. 2019a, b). This result is in line with previous findings (Ballard and Johnson 2004; Shavelson et al. 2019a, b) and supports the assumption that mathematics-specific aspects could contribute to knowledge and skills for teaching economics. Therefore, teacher knowledge and skills particularly in mathematics may foster the acquisition of teacher knowledge and skills in economics (in all three stages of teacher education).

In this study, not all relationships between all central constructs of our theoretical framework (Figure 1) could be empirically tested. Nonetheless, our findings significantly contribute to a more elaborate understanding of the relationships between domain-specific knowledge, generic attributes and instructional skills. In our domain-comparative study, we gain first empirical insights into relationship

patterns in two different domains. These findings can be considered first indications regarding the domain-specificity of teachers' instructional skills. The assessments developed and validated in our study can be used in future empirical research and expanded to further constructs and subjects.

References

- Aloe, A. M., & Becker B. J. (2009). Teacher verbal ability and school outcomes: Where is the evidence? *Educational Researcher*, 38, 612–624.
- Aydin, M. K., Bavli, B., & Alci, B. (2013). Examining the effects of pre-service Teachers' personality traits on their teaching competencies. *International Online Journal of Educational Sciences*, 5(3), 575–586.
- Baer, M., Dörr, G., Fraefel, U., Kocher, M., Küster, O., ...Wyss, C. (2007). Werden angehende Lehrpersonen durch das Studium kompetenter? Kompetenzaufbau und Standarderreichung der berufswissenschaftlichen Ausbildung an drei Pädagogischen Hochschulen in der Schweiz und in Deutschland. *Unterrichtswissenschaft*, 35(1), 15–47.
- Ballard, C. L., & Johnson, M. F. (2004). Basic math skills and performance in an introductory economics class. *The Journal of Economic Education*, 35(1), 3–23.
- Bastian, K. C., McCord, D. M., Marks, J. T., & Carpenter, D. (2017). A Temperament for teaching? Associations between personality traits and beginning teacher performance and retention. *AERA Open*, 3(1), 1–17. doi:10.1177/2332858416684764.
- Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9(4), 469–520.
- Baxter, J. A., & Lederman, N. G. (1999). Assessment and measurement of pedagogical content knowledge. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining Pedagogical Content Knowledge: The Construct and its Implications for Science Education* (pp. 147–161). Dordrecht: Kluwer Academic Publishers.
- Beck, E., Baer, M., Guldemann, T., Bischoff, S., Brühwiler, C., Müller, P. ...Vogt, F. (2008). *Adaptive Lehrkompetenz. Analyse und Struktur, Veränderbarkeit und Wirkung handlungssteuernden Lehrerwissens*. Münster: Waxmann.
- Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75, 729–750.
- Berliner, D. C. (1995). Teacher expertise. In B. Moon & A. S. Mayes (Eds.), *Teaching and Learning in the Secondary School* (pp. 46–52). London: Routledge.
- Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (2016). The relation between content-specific and general teacher knowledge and skills. *Teaching and Teacher Education*, 56, 35–46. <https://doi.org/10.1016/j.tate.2016.02.003>.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015a). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13. doi:10.1027/2151-2604/a000194.
- Blömeke, S., Hoth, J., Döhrmann, M., Busse, A., Kaiser, G., & König, J. (2015b). Teacher change during induction: Development of beginning primary teachers' knowledge, beliefs and performance. *International Journal of Science and Mathematics Education*, 13, 287–308. doi: 10.1007/s10763-015-9619-4.

- Blömeke, S., Baack, W., Dunekacke, S., Grassmann, M., Jenßen, L., Wedekind, H., ...Koinzer, T. (2014). *Effects of opportunities to learn on the mathematics pedagogical content knowledge of kindergarten teachers*. AERA 2014 Annual Meeting.
- Borko, H., & Shavelson, R. J. (1990). Teacher decision making. In B. F. Jones & L. Idol (Eds.), *Dimensions of Thinking and Cognitive Instruction* (pp. 311–346). Hillsdale, New Jersey: Lawrence Erlbaum.
- Bromme, R., & Haag, L. (2008). Forschung zur Lehrerpersönlichkeit. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (2nd ed., pp. 803–820). Wiesbaden: Verlag für Sozialwissenschaften.
- Bronkhorst, L. H., Meijer, P. C., Koster B., & Vermunt, J. D. (2014). Deliberate practice in teacher education. *European Journal of Teacher Education* 37(1): 18–34.
- Buschang, R. E., Chung, G. K. W. K., Delacruz, G. C., & Baker, E. L. (2012). Validating measures of algebra teacher subject matter knowledge and pedagogical content knowledge. *Educational Assessment* 17, Issue 1. doi: 10.1080/10627197.2012.697847.
- Carter, K. (1990). Teachers' knowledge and learning to teach. In W. R. Houston, M. Haberman & J. Sikula (Eds.), *The Handbook of Research on Teacher Education* (pp. 291–310). New York: Macmillan.
- Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of trainings motivation: A meta-analytic path analysis of 20 years research. *Journal of Applied Psychology*, 85(5), 678–707.
- Corcoran, R. P., & O'Flaherty, J. (2018). Factors that predict pre-service teachers' teaching performance. *Journal of Education for Teaching*, 44(2), 175–193. doi: 10.1080/02607476.2018.1433463.
- Corno, L., & Snow, R. (1986). Adapting teaching to individual differences among learners. In M. Wittrock (Ed.), *Handbook of Research on Teaching* (pp. 605–629). New York: Macmillan.
- Council for Economic Education. (2010). *Voluntary national content standards in economics* (2nd ed.). New York, NY: Council for Economic Education.
- Cutchin, G. C. (1998). *Relationships between the Big Five personality factors and performance criteria for in-service high-school teachers*. ProQuest dissertations and theses database. (UMI no. 9900173). West Lafayette, IN: Purdue University.
- Dalbert, C., & Radant, M. (2010). Ungewissheitstoleranz bei Lehrkräften. *Journal für LehrerInnenbildung*, 10(2), 53–57.
- Darling-Hammond, L., & Baratz-Snowden, J. (Eds.). (2005). *A good teacher in every classroom. Preparing the highly qualified teachers our children deserve*. San Francisco: Jossey-Bass.
- Darling-Hammond, L., & Lieberman, A. (2012). *Teacher education around the world: Changing policies and practices. Teacher quality and school development*. London: Routledge.
- Darling-Hammond, L., Newton, S. P., & Chung Wei, R. (2013). Developing and assessing beginning teacher effectiveness: the potential of performance assessments. *Educational Assessment, Evaluation and Accountability*, 25, 179–204. doi: 10.1007/s11092-013-9163-0.
- Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education*, 34, 12–25.
- Dreher, A., Lindmeier, A., Heinze, A., & Niemand, C. (2018). What kind of content knowledge do secondary mathematics teachers need? A conceptualization taking into account

- academic and school mathematics. *Journal Für Mathematik-Didaktik*, 39(2), 319–341. <https://doi.org/10.1007/s13138-018-0127-2>
- Ericsson, K. A. (2000). Expert performance and deliberate practice. <http://www.psy.fsu.edu/faculty/ericsson/ericsson.exp.perf.html>. Accessed: June 11, 2019.
- Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3), 363–406. doi: 10.1.1.169.9712.
- Eulenberger, J. (2015). Die Persönlichkeitsmerkmale von Personen im Kontext des Lehr-er_innenberufs. *SOE Papers on Multidisciplinary Panel Data Research*, 788, 1–21.
- Fenstermacher, G. D. (1994). The knower and the known: The nature of knowledge in research on teaching. *Review of Research in Education*, 20, 3–56.
- Gerlitz, J. Y., & Schupp, J. (2005). *Zur Erhebung der Big-Five-basierten Persönlichkeitsmerkmale im SOEP [The measurement of the Big Five personality traits in the SOEP]*. Berlin: DIW.
- Gess-Newsome, J. (1999). Pedagogical content knowledge. An introduction and orientation. In J. Gess-Newsome & N. G. Lederman (Eds.), *PCK and Science Education* (pp. 3–17). Dordrecht: Kluwer.
- Gess-Newsome, J., & Lederman, N. G. (1999). *Examining pedagogical content knowledge: The construct and its implications for science education*. Dordrecht: Kluwer Academic Publishers.
- Ghaith, G., & Yaghi, H. (1997). Relationships among experience, teacher efficacy, and attitudes toward the implementation of instructional innovation. *Teaching and Teacher Education*, 13(4), 451–458. doi: 10.1016/S0742-051X(96)00045-5.
- Gitomer, D., & Bell, C. (Eds.). (2016). *Handbook of research on teaching* (5th ed.). Washington, DC: American Educational Research Association.
- Guskey, T. R. (1988). Teacher efficacy, self-concept, and attitudes toward the implementation of instructional innovation. *Teaching and Teacher Education*, 4(1), 63–69. doi: 10.1016/0742-051x(88)90025-x.
- Hepberger, B., Moser Opitz, E., Heinze, A., & Lindmeier, A. (2019). Entwicklung und Validierung eines Tests zur Erfassung der mathematikspezifischen professionellen Kompetenzen von frühpädagogischen Fachkräften. *Psychologie in Erziehung und Unterricht*.
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42(2), 371–406. doi: 10.3102/00028312042002371.
- Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal*, 105(1), 11–30. doi: 10.1086/428763.
- Holtsch, D., Hartig, J., & Shavelson, R. (2018). Do practical and academic preparation paths lead to differential commercial teacher “quality”? *Vocations and Learning*, 1, 1–24. doi: 10.1007/s12186-018-9208-0.
- Holzberger, D., Philipp, A., & Kunter, M. (2014). Predicting teachers' instructional behaviors: The interplay between self-efficacy and intrinsic needs. *Contemporary Educational Psychology*, 39(2), 100–111.
- Holzberger, D., Philipp, A., & Kunter, M. (2013). How teachers' self-efficacy is related to instructional quality: A longitudinal analysis. *Journal of Educational Psychology*, 105(3), 774–786. doi: 10.1037/a0032198.
- Jackson, P. W. (1990). *Life in classrooms*. Teachers College Press.

- Jeschke, C., Kuhn, C., Lindmeier, A., Zlatkin-Troitschanskaia, O., Saas, H., & Heinze, A. (2019a). Performance assessment to investigate the domain-specificity of instructional skills among pre-service and in-service teachers of mathematics and economics. *British Journal of Educational Psychology*. doi: 10.1111/bjep.12277.
- Jeschke, C., Kuhn, C., Lindmeier, A., Zlatkin-Troitschanskaia, O., Saas, H., & Heinze, A. (2019b). What is the relationship between knowledge in mathematics and knowledge in economics? Investigating the professional knowledge of (prospective) teachers trained in two subjects. *Zeitschrift für Pädagogik*, 4, 511–524.
- Job, A. P. (2004). *The relationship between personality, occupation and student evaluations of teaching effectiveness*. ProQuest dissertations and theses database. (UMI no. 3127387). Portland, Oregon: Portland State University.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of Personality: Theory and Research* (pp. 114–158). New York, NY: Guilford Press.
- Kaiser, J., Helm, F., Retelsdorf, J., Südkamp, A., & Möller, J. (2012). Zum Zusammenhang von Intelligenz und Urteilsgenauigkeit bei der Beurteilung von Schülerleistungen im Simulierten Klassenraum. *Zeitschrift für Pädagogische Psychologie*, 4, 251–261.
- Keller-Schneider, M. (2009). Was beansprucht wen? Entwicklungsaufgaben von Lehrpersonen im Berufseinstieg und deren Zusammenhang mit Persönlichkeitsmerkmalen. *Unterrichtswissenschaft*, 37(2), 145–163.
- Kennedy, M. M., Ahn, S., & Choi, J. (2008). The value added by teacher education. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre & K. E. Demers (Eds.), *Handbook of Research on Teacher Education* (3rd ed., pp. 1247–1271). New York: Routledge.
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, 49(3), 568–589. doi:10.3102/0002831212437853.
- Kersting, N. B., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education*, 61(1–2), 172–181.
- Klassen, R. M., & Tze, V. M. C. (2014). Teachers' self-efficacy, personality, and teaching effectiveness: A meta-analysis. *Educational Research Review*, 12, 59–76.
- Klassen, R. M., Tze, V. M. C., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998–2009: Signs of progress of unfulfilled promise? *Educational Psychology Review*, 23(1), 21–43. doi: 10.1007/s10648–010-9141–8.
- Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., & Baumert, J. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education. *Journal of Teacher Education* 64(1), 90–106. doi: 10.1177/0022487112460398.
- Knievel, I., Lindmeier, A., & Heinze, A. (2015). Beyond knowledge: Measuring primary teachers' subject-specific competences in and for teaching mathematics with items based on video vignettes. *International Journal of Science and Mathematics Education*, 13(2), 309–329. doi: 10.1007/s10763–014-9608-z.
- König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A., & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education*, 38(1), 76–88.

- König, S. (2003). *Der Einfluss von Ungewissheitstoleranz auf den Umgang von Lehrenden mit schulischen Belastungen – eine quantitative Analyse an Berufsschulen*. Halle (Saale).
- König, S., & Dalbert, C. (2007). Ungewissheitstoleranz und der Umgang mit beruflich ungewissen Situationen im Lehramt. *Empirische Pädagogik*, 21(3), 306–321.
- Koşar, S. (2015). Trust in school principal and self-efficacy as predictors of teacher professionalism. *Eğitim ve Bilim*, 40(181), 255–270. doi: 10.15390/EB.2015.4562.
- Kuhn, C. (2014). *Fachdidaktisches Wissen von Lehrkräften im kaufmännisch-verwaltenden Bereich. Modellbasierte Testentwicklung und Validierung* [Pedagogical Content Knowledge of (Future) Teachers in Business and Economics: Theoretical Modeling, Test Development, and Validation]. Landau: Verlag Empirische Pädagogik.
- Kuhn, C., Zlatkin-Troitschanskaia, O., Brückner, S., & Saas, H. (2018). A new video-based tool to enhance teaching economics. *International Review of Economics Education*, 27, 24–33. doi: 10.1016/j.iree.2018.01.007.
- Kuhn, C., Alonzo, A. C., & Zlatkin-Troitschanskaia, O. (2016). Evaluating the pedagogical content knowledge of pre- and in-service teachers of business and economics to ensure quality of classroom practice in vocational education and training. *Empirical Research in Vocational Education and Training*, 8(1). doi: 10.1186/s40461-016-0031-2.
- Kuhn, C., Happ, R., Zlatkin-Troitschanskaia, O., Beck, K., Förster, M., & Preuße, D. (2014). Kompetenzentwicklung angehender Lehrkräfte im kaufmännisch-verwaltenden Bereich – Erfassung und Zusammenhänge von Fachwissen und fachdidaktischem Wissen. In E. Winther & M. Prenzel (Eds.), *Perspektiven der empirischen Berufsbildungsforschung: Kompetenz und Professionalisierung. Zeitschrift für Erziehungswissenschaft*, 17(1), 149–167. doi:10.1007/s11618-013-0456-3.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance. Can one construct predict them all? *Journal of Personality and Social Psychology*, 86(1), 148–161.
- Kunina-Habenicht, O., Maurer, C., Schulze-Stocker, F., Wolf, K., Hein, N., Leutner, D., ... Kunter, M. (2019). Zur curricularen Validität des BilWiss 2.0-Tests zur Erfassung des bildungswissenschaftlichen Wissens von (angehenden) Lehrkräften. *Zeitschrift für Pädagogik*, 4, 542–556. doi: 10.3262/ZP1904542.
- Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2011): *Die professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms CO-ACTIV*. Münster: Waxmann.
- Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R*. Göttingen: Hogrefe.
- Lin, X., Schwartz, D. L., & Hatano, G. (2005). Toward teacher's adaptive metacognition. *Educational Psychologist*, 40(4), 245–255.
- Lindmeier, A. (2011). *Modeling and measuring knowledge and competencies of teachers: A threefold domain-specific structure model, exemplified for mathematics teachers, operationalized with computer- and video-based methods*. Münster: Waxmann.
- Lindmeier, A., Seemann, S., Wullschleger, A., Meier, A., Dunekacke, S., Moser Opitz, E., Heinze, A., Leuchter, M., & Vogt, F. (under review). Early childhood teachers' domain-specific professional competences and their relation to the quality of mathematical learning situations – An aspect of predictive validity. *ZDM*.

- Llinares, S., & Krainer, K. (2006). Mathematics (student) teachers and teacher educators as learners. In A. Gutiérrez & P. Boero (Eds.), *Handbook of Research on the Psychology of Mathematics Education* (pp. 429–460). Roderdam: Sense Publishers.
- Loch, C., Lindmeier, A. M., & Heinze, A. (2015). The Missing Link? – School-Related Content Knowledge of Pre-Service Mathematics Teachers. In K. Beswick, T. Muir, & J. Fielding-Wells (Eds.), *Proceedings of the 39th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 209–216). Hobart: PME.
- Lohmann, M., & Andere (1966). *Certain characteristics of student teachers who stay in teaching*. <https://eric.ed.gov/?id=ED010009>. Accessed: June 11, 2019.
- Mankiw, N. G. (2012). *Principles of Economics* (6th Ed.). Mason, OH: South-Western.
- Mathesius, S., Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2019). Überprüfung eines Tests zum wissenschaftlichen Denken unter Berücksichtigung des Validitätskriteriums relations-to-other-variables. *Zeitschrift für Pädagogik*, 4, 492–510. doi: 10.3262/ZP1904492.
- Mayr, J. (2016). Lehrerpersönlichkeit. In M. Rothland (Eds.), *Beruf Lehrer. Ein Studienbuch* (pp. 87–102). Münster: Waxmann.
- Mayr, J. (2011). Der Persönlichkeitsansatz in der Forschung zum Lehrberuf. Konzepte, Befunde und Folgerungen. In E. Terhart, H. Bennewitz & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrberuf* (pp. 189–215). Münster: Waxmann.
- Morris-Rothschild, B. K., & Brassard, M. R. (2006). Teachers' conflict management styles: The role of attachment styles and classroom management efficacy. *Journal of School Psychology*, 44(2), 105–121. doi: 10.1016/j.jsp.2006.01.004.
- Niermann, A. (2017). *Professionswissen von Lehrerinnen und Lehrern des Mathematik- und Sachunterrichts*. '...man muss schon von der Sache wissen.'. <http://nbn-resolving.de/urn:nbn:de:0111-pedocs-125876>. Accessed: June 11, 2019.
- Nilsson, P. (2014). When teaching makes a difference: Developing science teachers' pedagogical content knowledge through learning study. *International Journal of Science Education* 36 (11), 1794–1814. doi: 10.1080/09500693.2013.879621.
- Nilsson, P. (2008). Primary science student teachers' and their mentors' joint learning through reflection on their science teaching. *Journal of Science Teacher Education*. <http://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-3652>. Accessed: June 11, 2019.
- Nilsson, P., & Loughran, J. (2012). Exploring the development of pre-service science elementary teachers' pedagogical content knowledge. *Journal of Science Teacher Education* 23(7), 699–721. doi:10.1007/s10972-011-9239-y.
- Oser, F., Salzmann, P., & Heinzer, S. (2009). Measuring the competence-quality of vocational teachers: An advocacy approach. *Empirical Research in Vocational Education and Training*, 1(1), 65–83.
- Praetorius, A.-K., Drexler, K., Rösch, L., Christophel, E., Heyne, N., Scheunpflug, A., ... Dresel, M. (2015). Judging students' self-concepts within 30 seconds? An application of the zero-acquaintance approach to research on teachers' judgment accuracy. *Learning and Individual Differences*, 37, 231–236.
- Raven, J. C. (1962). *Advanced Progressive Matrices*. London: Lewis & Co. Ltd.
- Reis, J. (1996). *Inventar zur Messung der Ambiguitätstoleranz (IMA)*. Heidelberg: Roland Asanger.
- Richmond, G., Salazar, M. d. C., & Jones, N. (2019). Assessment and the future of teacher education. *Journal of Teacher Education*. doi: 10.1177/0022487118824331.

- Riese, J., & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen. *Zeitschrift für Erziehungswissenschaft*, *15*(1), 111–143.
- Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy* *6*(1), 43–74.
- Ross, J. A. (1998). The antecedents and consequences of teacher efficacy. In J. Brophy (Ed.), *Advances in Research on Teaching* (pp. 385–400). Greenwich, CT, US: JAI.
- Sadler, D. R. (2013). Making competent judgements of competence. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.) *Modeling and Measuring Competencies in Higher Education* (pp. 13–27). Heidelberg: Springer.
- Santagata, R., & Sandholtz, J. H. (2018). Preservice teachers' mathematics teaching competence: Comparing performance on two measures. *Journal of Teacher Education*, doi:10.1177/0022487117753575.
- Santagata, R., & Yeh, C. (2016). The role of perception, interpretation, and decision making in the development of beginning teachers' competence. *ZDM Mathematics Education*, *48*(1), 153–165. doi: 10.1007/s11858–015-0737–9.
- Schmitz, G. S., & Schwarzer, R. (2000). Perceived self-efficacy of teachers: Longitudinal findings with a new instrument. *German Journal of Educational Psychology*, *14*(1), 12–25. doi: 10.1024//1010–0652.14.1.12.
- Schönfeld, A. H., & Kilpatrick, J. (2008). Toward a theory of proficiency in teaching mathematics. In D. Tirosh & T. Wood (Eds.), *International Handbook of Mathematics Teacher Education* (2nd ed., pp. 321–354). Rotterdam: Sense Publishers.
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.
- Schwarzer, R., & Jerusalem, M. (2002). Das Konzept der Selbstwirksamkeit. *Zeitschrift für Pädagogik Beiheft*, *44*, 28–53.
- Schwarzer, R., & Warner, L. M. (2014). Forschung zur Selbstwirksamkeit bei Lehrerinnen und Lehrern. In E. Terhart, H. Bennewitz & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* (2nd ed., pp. 662–678). Münster/New York: Waxmann.
- Seidel, T., & Stürmer, K. (2014). Modeling and measuring the structure of professional vision in preservice teachers. *American Educational Research Journal*, *51*(4), 739–771.
- Seifried, J., & Wuttke, E. (2015). Was wissen und können (angehende) Lehrkräfte an kaufmännischen Schulen? Empirische Befunde zur Modellierung und Messung der professionellen Kompetenz von Lehrkräften. In S. Schumann (Ed.), *Ökonomische Kompetenzen in Schule, Ausbildung und Hochschule*. (1st ed., pp. 125–145). Landau/Pfalz: Empirische Pädagogik.
- Shavelson, R. J., Marino, J., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2019a). Reflections on the assessment of quantitative reasoning. In B.L. Madison, & L.A. Steen (Eds.), *Calculation va. context: Quantitative literacy and its implications for teacher education* (2nd ed.). Washington, D.C.: Mathematical Association of America.
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. P. (2019b). Assessment of university students' critical thinking: Next generation performance assessment. *International Journal of Testing*. doi: 10.1080/15305058.2018.1543309.
- Short, G. (1995). Understanding domain knowledge for teaching: Higher-order thinking in pre-service art teacher specialists. *Studies in Art Education*. *36*(3), 154–169. doi: 10.1080/00393541.1995.11649975.

- Shulman, L. S. (1986a). Paradigms and research programs in the study of teaching: a contemporary perspective. In M.C. Wittrock (Ed.), *Handbook of research on teaching*. New York: MacMillan.
- Shulman, L. S. (1986b). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14.
- Sorrentino, R. M., Short, J.-A. C., & Raynor, J. O. (1984). Uncertainty orientation: Implications for affective and cognitive views of achievement behavior. *Journal of Personality and Social Psychology*, 46(1), 189–206. doi: 10.1037/0022–3514.46.1.189.
- Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17, 783–805.
- Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, 68(2), 202–248. doi: 10.3102/00346543068002202.
- Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and Selecting Key Competencies* (pp. 45–65). Seattle: Hogrefe & Huber.
- Wolters, C. A., & Daugherty, S. G. (2007). Goal structures and teachers' sense of efficacy: Their relation and association to teaching experience and academic level. *Journal of Educational Psychology*, 99(1), 181–193. doi: 10.1037/0022–0663.99.1.181.
- Woolfolk, A. E., Rosoff, B., & Hoy, W. K. (1990). Teachers' sense of efficacy and their beliefs about managing students. *Teaching and Teacher Education*, 6(2), 137–148. doi: 10.1016/0742–051x(90)90031-y.
- Wyss, C., Kocher, M., & Baer, M. (2013). Erwerb und Erfassung unterrichtlicher Kompetenzen im Lehrstudium und im Übergang in den Beruf. Ein multiperspektivischer Ansatz zur Wirksamkeit der Ausbildung und der Auswirkung der Berufspraxis. In U. Riegel & M. Klaas (Eds.), *Videobasierte Kompetenzforschung in den Fachdidaktiken* (pp. 283–301). Münster: Waxmann.
- Zlatkin-Troitschanskaia, O., & Pant, H. A. (2016). Measurement advances and challenges in competency assessment in higher education. *Journal of Educational Measurement*, 53(3), 253–264. doi: 10.1111/jedm.12118.
- Zlatkin-Troitschanskaia, O., Kuhn, C., Brückner, S., & Leighton, J. P. (2019a). Evaluating a technology-based assessment (TBA) to measure teachers' action-related and reflective skills. *International Journal of Testing (IJT)*, 19(2), 148–171. doi: 10.1080/15305058.2019.1586377.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019b). On the complementarity of holistic and analytic approaches to performance assessment scoring. *The British Journal of Educational Psychology*. Advance online publication. doi: 10.1111/bjep.12286.
- Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Ed.), *Higher education learning outcomes assessment – International perspectives* (pp. 175–197). Frankfurt/Main: Lang.
- Zumwalt, K., & Craig, E. (2005). Teachers' characteristics: Research on the indicators of quality. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 157–260). Mahwah, N. J.: Erlbaum.



2.5 Development of Prospective Physics Teachers' Professional Knowledge and Skills during a One-Semester School Internship

Vogelsang, C., Borowski, A., Kugelmeyer, C., Riese, J., Buschhüter, D., Enkrott, P., Kempin, M., Reinhold, P., Schecker, H., and Schröder, J.

Abstract

In academic teacher education programs prospective physics teachers are supposed to acquire professional knowledge and skills that enable them to carry out effective instruction. However, it is unclear which knowledge has an impact on teaching quality and how knowledge and teaching skills develop throughout studies. In particular, this is the case for practical parts of teacher education programs like school internships or other forms of field experiences. Therefore, we examine the development of pre-service physics teachers' professional knowledge over a one-semester internship in schools. Furthermore, we analyze the development of their skills in (a) planning physics lessons, (b) reflecting physics lessons, and (c) explaining physics over that internship using an innovative approach of standardized performance assessments. So far, our longitudinal analyses of a cohort of prospective physics teachers from four German universities hardly show any development of their professional knowledge and skills during the internship. Further analyses are needed to gain more insight into this low efficacy of the internship.

Keywords

Science education, pre-service teacher education, professional knowledge, teachers' professional development, performance assessment, pre-post-measurement, field experience

1 Introduction

The primary goal of teacher education is to support future teachers in developing professional knowledge and skills to meet the challenges of their profession. Common models treat professional knowledge as a key component of professional competence, whereas competence can be described as 'the latent cognitive and affective-motivational underpinning of domain-specific performance in varying situations' (Blömeke et al. 2015, p. 3). Therefore, a great amount of teacher education programs focus on the development of such knowledge, especially in the German teacher education system. In Germany, teacher education consists of three consecutive phases necessary to become a teacher (Cortina and Thames 2013). In the first phase, future teachers enrol in a teacher education study program at a university, including a three-year bachelor's degree followed by a two-year master's degree. The second phase consists of a 12 to 24 months in-school induction program. While the first phase focuses on the acquisition and development of theoretical knowledge, the second phase emphasizes practical teacher training. The third phase aims at the further professional development of in-service teachers. The underlying model of these phases can be described as a functional chain (Diez 2010). During their university studies, future teachers are meant to acquire knowledge, which they have to apply in their teacher training course afterwards. In this perspective, it is assumed that teachers use their professional knowledge as a resource to perform their daily tasks. This model poses some challenges for teacher educators creating teacher education programs. They have to ensure that prospective teachers acquire knowledge, which is relevant for teaching. It has to be part of the knowledge base for teaching (van Driel et al. 2001), which allows teachers to develop skills to carry out high-quality instruction. To meet this challenge, most federal states in Germany implemented a one-semester internship at a school as part of their master's degree programs for teachers (practical semester). It is meant to enable the use of theoretical knowledge and to gather first teaching experiences already before the second phase. During their one-semester internship, in addition to teacher training by expert teachers at school, all participating students also attend supporting courses at the university (usually one day of the

week). However, it remains an open question to which extent academic teacher education programs contribute to the development of professional knowledge and skills and which types of knowledge have an impact on the quality of performance in teaching situations. This is especially the case for field experiences in teacher education like long-term internships at schools.

In our research project Profile-P+ (*Professional Knowledge in Academic Physics Teacher Education*), these questions are addressed. We focus on prospective physics teachers for secondary schools in Germany. We examine the development of their professional knowledge by longitudinal studies over the first two years of their bachelor's degree program and evaluate their development of professional knowledge and skills to cope with three typical requirements for physics teachers (planning physics lessons, reflecting on physics teaching, explaining physics) in a longitudinal section during a practical semester. Moreover, we analyze the relationship between professional knowledge and skills. In this chapter, we present preliminary results regarding professional development during the one-semester internship.

2 Theoretical Background

2.1 Physics Teachers' Professional Knowledge

Blömeke, Gustafson and Shavelson (2015) describe competence as a continuum that regards cognitive and affective-motivational dispositions as the basis for situation-specific skills, which in turn enable performance in complex real-life situations. Professional knowledge is seen as one key-part of these dispositions of the professional competence for teaching. Following recent work based on the influential considerations of Shulman (1986), we focus on three dimensions of professional knowledge: content knowledge, pedagogical content knowledge and general pedagogical knowledge. Content knowledge (CK), also known as subject matter knowledge, comprises knowledge of the contents and methods of the subject taught as well as epistemological aspects. For the domain of physics, there are different models of content knowledge (Woitkowski and Borowski 2017), which usually differentiate the "depth" of knowledge according to curricular levels (school knowledge, university knowledge). We extend this structure by a specific form of CK assumed to be a specific basis for teaching: deeper school knowledge. In our approach, deeper school knowledge refers to a meta-perspective on school knowledge, for instance, identifying suitable problem solving strategies for specific physics problems or the discussing structural relationships between physics

concepts. Pedagogical content knowledge (PCK) refers to knowledge that teachers need to prepare and structure content in a way that is appropriate for their students. For the domain of science (especially physics), a large number of models have been proposed, which have recently been brought together in the Refined Consensus model of PCK (Carlson et al. 2019). For example, PCK for physics teaching contains knowledge about students' alternative ideas of physics concepts, about instructional strategies and about implementing experiments in physics instruction. Finally, pedagogical knowledge (PK) contains knowledge about learning principles and pedagogical concepts like knowledge about classroom management or motivation (e.g. König et al. 2011). Surely, CK, PCK, and PK do not cover the full complexity of teachers' professional knowledge; however, these three dimensions are mirrored in the typical structure of academic physics teachers education in Germany. During their studies, student physics teachers take courses focusing on content knowledge (CK), general pedagogy (PK) and also on concepts of physics education (PCK).

2.2 The Influence of Physics Teachers' Professional Knowledge on Their Teaching Skills

The professional skills of physics teachers have been assessed by the quality of their instruction (Diez 2010). However, little correlation has been observed between physics-related professional knowledge and the quality of physics teaching or student learning in physics lessons. Cauet et al. (2015) could not find any correlation between physics teachers' professional knowledge (CK and PCK measured by written tests containing multiple choice and open questions), students learning (measured by written tests containing multiple choice and open questions) and the level of cognitive activation of the observed physics lessons (measured by video-analysis) in German secondary schools. Similar results are reported by Ohle, Boone & Fischer (2014). They analyzed the impact of K4-elementary teachers' CK of a specific physics topic on their students' achievement (both measured by multiple choice tests). „Results showed that neither teachers' interest nor content knowledge impacted students' outcomes directly.“ (Ohle et al. 2014, p. 14).

These results are independent of how quality of teaching was measured, for instance, by high inference ratings of instruction (e.g. Korneck et al. 2017) or by analyzing the content structure of observed lessons (Liepertz and Borowski 2018). In the sense of the Refined Consensus Model, hardly any connections between enacted PCK and personal or collective PCK could be observed (Carlson et al. 2019). This might indicate a lack of relevance of the assessed knowledge for teaching.

However, in all of these studies, the assessment instruments used show good curricular validity often based on expert ratings of teacher educators or experienced physics teachers.

Another reason for the missing link between teaching quality and students' achievement might result from the complexity of real teaching situations. Teaching in real classrooms is affected by many heterogeneous context factors that might suppress existing relations between teachers' knowledge and their actual teaching performance (Kulgemeyer and Riese 2018). To obtain a clearer picture, the quality of teaching should, therefore, be researched in a more standardized way. This can be achieved according to concepts from medicine education (Miller 1990). So-called 'Objective Structured Clinical Examinations (OSCE)' simulate typical standard situations for health professionals, often with the help of trained actors. These scenarios are designed as authentic as possible to mirror real situations occurring in everyday professional life. At the same time, they are standardised in such a way that a high degree of comparability is guaranteed. Scoring sheets can be used to achieve high test standards (Walters et al. 2005). Kulgemeyer and Riese (2018) developed such a performance assessment for explaining physics in the form of a role-playing situation. They examined the correlation between prospective physics teachers' professional knowledge (CK and PCK, measured by written tests containing multiple choice and open questions) and their skills in explaining physics (measured by performance assessment). The results showed significant correlations between explaining skills and professional knowledge (for details, see Kulgemeyer and Riese 2018, p. 18).

2.3 Development of Professional Knowledge and Skills during Field Experiences

School internships or field experiences are an integral component of typical teacher education programs. Internships have multiple aims, for example, to provide opportunities for prospective teachers to make first teaching experiences on their own, to acquire basic skills for teaching (like planning a lesson or reflect on one's teaching) and to cope with the gap between academic learning and the demands of professional practice in the field (Cohen et al. 2013). However, it is unclear, whether internships contribute to those aims. Most research on the effectiveness of internships or field experiences, especially in the context of German teacher education, relies on student teachers' self-reports like self-rated skills or self-rated instructional quality (Besa and Büdcher 2014). Self-reports are often suspected of bias, like the tendency of participants to rate themselves in a favourably way. How-

ever, studies hardly used more proximal assessments to evaluate the development of skills during internships. Holtz and Gnambs (2017) examined the change of instructional quality of student teachers throughout a 15-week school internship, measured by ratings of two observed lessons by different groups, expert teachers, students, and self-ratings. The results showed an improvement of instructional quality, but they are related to many different school subjects not specifically to physics. Volmer et. al (2019) investigated the skills of prospective elementary teachers during a one-semester internship using open written reflections on a given videotaped lesson. They report a significant increase in reflection quality between before and after the internship.

3 Research Questions

Prior studies showed few and inconclusive results regarding the development of prospective physics teachers' professional knowledge during academic teacher education and the relation between knowledge and skills. To contribute to these research gaps, we address the following overarching research questions (RQ) in the project Profile-P+:

1. How does the professional knowledge (CK, PCK, PK) of prospective physics teachers develop over a one-semester internship (pre, post)?
2. How do the skills of *planning* physics lessons, *explaining* physics and *reflecting* on physics lessons develop over a one-semester internship (pre, post)?

In our study, we collect data of additional personal characteristics adjusted for mathematical skills, attitudes to explanation, learning opportunities, and demographic data.

Since there are hardly any opportunities for systematically learning more theoretical knowledge over a one-semester internship, we expect little to no increase in professional knowledge (RQ1). However, we expect substantial increases in all three analyzed skills (RQ2) as the internship should provide many opportunities for lesson planning and reflecting and explaining physics in classroom instruction in an authentic school setting.

4 Design and Sample

We used a longitudinal approach to follow cohorts of prospective physics teachers in master's degree programs at four German universities in three different federal states.

Although the internships vary in some details between the four participating universities, the overall structure is similar. During the internships – lasting about five months – the prospective physics teachers receive guidance in their teaching from experienced teacher mentors at a school four days a week. This part of the internship is supposed to provide practical insights into German teachers' daily routines and to enable first self-sufficient teaching experiences. The mentors are supposed to provide feedback on the instruction carried out by the teacher students and to help them reflect on their teaching. One day a week, the prospective physics teachers visit university courses (one for each of the two teaching subjects, one for general pedagogics). Part of these courses is a reflection on experiences made at the internship schools taking based on theoretical concepts of PCK and PK. Also, in these courses prospective physics teachers are supported in planning their lessons. This structure (in-school-training combined with university courses) is typical for one-semester internships in teacher education programs of the most federal states in Germany.

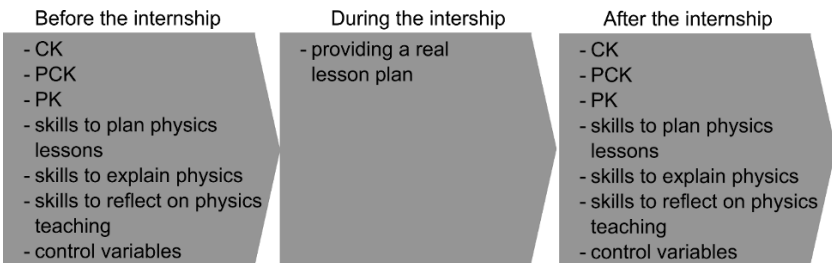


Figure 1 Study design: longitudinal section during the practical semester in the master's program

For our study, the professional knowledge (CK, PCK, PK), as well as the skills to plan physics lessons, to explain physics, and to reflect on physics teaching, were assessed before and after completing the one-semester internship (Figure 1). The extensive data collection required five test dates before and after the internship each, which were realised within the accompanying university courses taken by the students. Tests were scheduled on different days for each measurement point.

The total test time per person and measurement point was 330 minutes. The complete sample consists of $N = 80$ prospective physics teachers. On average, they were in the ninth semester of their study programmes (semester: $M = 9.03$; $SD = 3.03$) during the internship and 25 years old (age: $M = 25.41$; $SD = 5.09$). 61 % of the students are male, 39 % female.

The data of all participating universities has been pooled for analyses. Cohort effects have been checked. Due to the high testing time and the facts that the participation was voluntary not all teacher students took part in all the assessments. Therefore, we used different subsamples for analyses depending on the construct we focused on. To control for possible effects of decreasing test motivation we used a short test-motivation scale after every performance assessments (for details, see Vogelsang et al. 2019). Dependent t-tests were carried out to analyse changes of professional knowledge and performance during the internship (pre, post).

5 Instruments

For data collection, we used two widely tested and comprehensively validated instruments from our previous project Profile-P (written test for PCK, performance assessment for the skills of explaining physics) (Riese et al. 2015). The instrument for the assessment of CK was adopted from Profile-P and further developed with greater emphasis on deeper school knowledge. The performance assessments for the skills of planning physics lessons and reflecting on physics lessons have been newly developed for this study. All the instruments refer to the physical content of mechanics to establish the necessary content comparability.

5.1 Content Knowledge

CK is seen as a major component of teachers' professional knowledge. In line with previous research we assume, that physics teacher need specific forms of CK, which is of special importance to plan high-quality physics lesson or to reflect on their teaching properly. Following already developed models (Kirschner 2013), we distinguish between three dimensions of CK: school-, university- and deeper school knowledge (SK, UK, DSK). School knowledge is described by the official school curricula and university knowledge can be operationalized by the university curriculum. Based on the approach of Riese et. al. (2015), we defined deeper school knowledge as (1) identifying relations between physics concepts, (2) handling model limitations, and (3) identifying suitable problem-solving approaches.

As described by Riese and Reinhold (2012), this knowledge is assumed to be of special importance for teachers.

Based on this model, we developed a written test consisting of 48 single-choice-items focusing on different aspects of mechanics (velocity and acceleration, Newton's Laws, Conservation of Energy). The total test time is 50 minutes. Curricular validity was ensured by analyses of typical school and university textbooks, by analyses of school curricula and analyses of the physics curricula of the participating universities. Stimulated recall interviews with $N = 8$ physics teacher students after taking the test indicate, that the items are perceived as appropriate tasks in academic teacher education. For construct validity, we investigated whether the structure of our model corresponds to the structure of the empirical data. We compared Rasch models with different dimensions based on data of $N = 861$ physics teacher students in a bachelor or master program at twelve German universities combined with students studying the research-oriented physics bachelor program. The results show that a three-dimensional model is to be preferred over two- or one-dimensional models. For all the sub-scales, satisfactory reliability was found (EAP-reliability ranging from 0.76 to 0.84; for details, see Vogelsang et al. 2019, p. 484).

5.2 Pedagogical Content Knowledge

For the assessment of PCK, another key component of physics teachers' professional knowledge, we used a written instrument already developed in our previous project Profile-P (Riese et al. 2015). The underlying comprehensive model of physics teachers' PCK was developed using different conceptualizations of PCK in science subjects (e.g. Lee & Luft 2008; Magnusson, Krajcik et al. 1999) and also considering curricula analyses. The test instrument focuses on four aspects of PCK: experiments, instructional strategies, students' misconceptions and how to deal with them and physics education concepts like conceptual change. It includes open situational judgment items as well as complex multiple-choice items (multiple select, 43 items). The total test time is 65 minutes. To ensure validity, several steps of validation were taken in our previous project Profile-P (Riese et al. 2015). We investigated content validity by analyzing curricula and expert ratings from educators at four universities. Also, a think-aloud study with $N = 15$ prospective physics teachers was carried out to check, if items could be solved using only CK rather than PCK (Gramzow et al. 2013). For construct validity, one-dimensional and four-dimensional Rasch models were compared, indicating a better matching of the data by a four-dimensional model (for details, see Kulgemeyer and Riese

2018). For all the sub-scales, satisfactory moderate reliability was found in different studies (e.g. Kugelmeyer and Riese 2018; Riese et al. 2015). For the rather small sample of this study we used manifest scores and found a rather low but sufficient reliability for the total score (*Cronbach's* $\alpha = 0.66$).

5.3 Pedagogical Knowledge

For the assessment of pedagogical knowledge, we used an adapted short-version (Riese and Reinhold 2012) of a written instrument from Seifert and Schaper (2012). The full version was used in several studies in Germany (Mertens and Gräsel 2018), also the short-version (Riese and Reinhold 2012). The short-version addresses two aspects of PK: general instructional strategies and classroom management. The total test time is 15 minutes. Since we focus on the subject-specific components of professional knowledge in Profile-P+, no further validations were carried out by ourselves. We rely on the results from previous research where sufficient arguments for construct validity and content validity regarding teacher education in Germany has been reported (for details, see Seifert and Schaper 2012).

5.4 Skills of Planning Physics Lessons

The process of teachers' lesson planning can be described as a recursive process. To plan a lesson, teachers have to analyze preconditions, plan certain classroom actions and reflect on their planning decisions (Shavelson and Stern 1981). Experienced teachers mostly do not elaborate on their lesson plans, as they have scripts and routines to fall back on (Stender 2014). However, beginning or student teachers need to develop such scripts. Therefore, they need to plan actual lessons. In doing so, it is assumed, that student teachers heavily rely on their – more theoretical – professional knowledge. To analyze this assumption, we developed a performance assessment for prospective physics teachers' skills to plan physics lessons. We followed Miller's (1990) approach and developed an instrument to assess planning skills in a standardized performance situation, in which student teachers need to plan a whole lesson instead of reproducing knowledge about lesson planning. The paper-pencil instrument puts students in a situation where they have to plan a lesson about Newton's third law. Therefore, a short description of the class and their learning prerequisites is provided and specific learning objectives are set. The lesson plan has to be documented on a prestructured planning paper, which suggests some mandatory parts of physics lessons to allow higher comparability.

To evaluate the quality of lesson plans, we developed a theoretical model, which contains different aspects of physics lesson planning (e.g. implementation of experiments, exercises, contexts, learning objectives and preconditions). The model was developed by using subject-specific literature about lesson planning and it was combined with an inductive approach, resulting in a codebook with currently $N = 59$ coding items. By using the codebook, we have so far coded $N = 141$ out of 160 individual lesson plans, resulting in 66 sets of pre and post data. For interrater agreement, $N = 52$ of the lesson plans were double coded. The agreement amounts to 89%, Gwet's AC 1 is .849, which indicates good agreement. We carried out multiple steps of validation. Interviews with three teacher trainers were conducted and their judgements regarding the quality of three selected lesson plans were compared to corresponding results provided by the codebook. The results indicate an agreement among the teacher trainers about the perceived order of quality of those lesson plans as well as an agreement of their mean grades' order to the order of the planning score provided by the codebook. The standardized lesson plans were also compared to real lesson plans, created during the internship by the same students. We found an indication of similar planning behaviour comparing the assessment and the real lesson plans. Based on all 59 coding-items a sum score was built. For the sample of this study, the scale-reliability of the total planning score is sufficient (*Cronbach's* $\alpha = 0.79$).

5.5 Skills of Explaining Physics

To assess the skills to explain physics, we used a performance assessment already developed in our previous project Profile-P (Riese et al. 2015). Explaining is core part of physics instruction and therefore, a central skill of physics teachers (Geelan 2012). Explanations in the context of instruction can be described as a dialogic process, in which the teacher tries to communicate a scientific concept to one or more students. In this process, the teacher has to consider two aspects. First, the explanation has to represent the scientific concept in an adequate way, for example, its structure, highlighting the major aspects (subject-adequate). Second, the teacher has to consider the students' needs, for example, considering their supposed prior knowledge or any misconceptions (addressee-oriented). During the process, the teacher also has to evaluate his explanation, for instance, by asking questions to the student. We developed a dialogic explaining performance assessment simulating an authentic face-to-face explaining situation. After a short preparation time using standardized materials, participants had to explain a topic of mechanics to a student. The explaining attempts were videotaped. The student has been trained

to behave in a standardized way during the explaining situation (e.g. giving specific prompts as feedback). All videotaped explanations were analyzed using the model of explaining physics by Kulgemeyer and Tomczyszyn (2015). This model distinguishes between the two aspects of explanation quality (subject-orientation, addressee-orientation) represented by 12 categories for appropriate resp. inappropriate explaining, for instance, explaining physics concepts in everyday language. Based on the codings of the videotaped explanations, an explaining performance Index (PI) was build. For this assessment, also several steps of validation have been conducted (Kulgemeyer and Riese 2018). The PI predicted expert decisions on for the better explaining quality when a pair of videos was compared to a moderate to great extent. In addition, expert interviews have been carried out to ensure content validity. Interrater-reliability of two independent raters reached accordance of 91 %. In previous studies, also suitable reliability of the PI has been reported (for details, see Kulgemeyer & Riese, 2018). For the sample of this study, the scale-reliability of the PI is sufficient (*Cornbach's* $\alpha = 0.77$).

5.6 Skills of Reflecting on Physics Lessons

Reflection can be described as a spontaneous, common, real thinking process that gives coherence to an initially incoherent and unclear situation (Clarà 2015). Plöger, Scholl and Seifert (2015) developed a multi-stage model for reflection on lessons, which was adapted for reflecting physics lessons by modifying specific challenges in the fields of CK and PCK (Figure 3). The model distinguishes three dimensions of reflection. The elements of reflection includes four steps of reflection: (1) description of the framework conditions of a lesson (e.g. students' pre-knowledge, teaching goals) and the teaching situation, (2) evaluation of the described teaching situation, (3) providing alternatives for the observed behaviour and (4) drawing of consequences for (a) the following lesson, (b) the development of the reflective individual, or (c), in case of reflecting others' actions, the development of the teacher observed (Nowak et al. 2019). The reasoning dimension represents another aspect of reflection quality. Reflection quality is higher, if evaluations, alternatives and consequences are reasoned rather than given as spontaneous subjective judgements. The last dimension indicates which knowledge base (CK, PCK, PK) is subject in the point of reflection addressed. For example, reflecting on how to deal with students' misconceptions would be based on PCK.

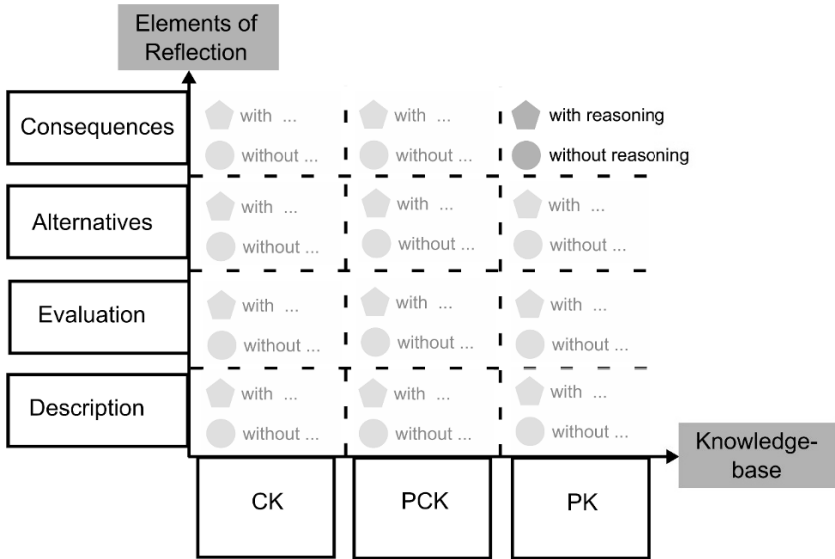


Figure 2 Model of reflections on physics lessons

Following Miller's (1990) approach, we developed a performance assessment for the skills of prospective physics teachers to reflect on physics lessons. We conduct an interactive online test simulating a situation, in which the student teachers have to reflect on the video-taped physics lesson of a fictive fellow intern. The fictive intern (Robert) asks the test person in a simulated dialogue for advice. The videotaped physics lesson focuses on Newton's third law and the conservation of momentum. It was created based on a script by persons acting as school students. The fictive lesson includes standard parts of physics instruction (e.g. an experiment) and typical problems of teaching physics (e.g. misconceptions) that provide reflection causes (RC). These RCs were assigned to the knowledge bases of the reflection model by an expert rating. During the test, the student teachers receive information about the context and they give Robert feedback on particular parts of the lessons. The verbal answers are recorded and categorised using qualitative content analysis by Mayring (2015).

The coding categories are used to evaluate the quality of reflection, for example, whether or not they give a reasoned evaluation or present alternatives for one RC. The sum of coded categories is assumed to be a measure for the skills to reflect on physics teaching. We formed a score for general reflection skills (RS). $N = 32$ com-

plete data-sets (pre and post) have already been analyzed. Interrater agreement has been determined by triple-coding of 17 % of the performance assessments reaching an average of .882 for Gwet's AC1. That indicates a very good agreement. For the small (sub)sample of this study the scale reliability of the reflection score is rather low (*Cronbach's* $\alpha = 0.46$) before resp. after (*Cronbach's* $\alpha = 0.52$) the internship so far.

6 Results

Data of between 20 and 80 physics teacher students has been analyzed so far, depending on the particular construct. In this chapter, we present preliminary results.

Tables 1 and 2 describe the results regarding the development of professional knowledge differentiated into subscales for each test. CK was scaled using a three-dimensional Rasch model (for the scoring, see Section 5.1). Therefore, latent EAP-scores are reported. For PCK and PK manifest sum-scores are reported. All analyses result from dependent t-tests for particular knowledge scores.

Table 1 Development of CK (latent scores)

	N	t_1		t_2			
CK	80	M	SD	M	SD	p	d
school knowledge (SK)		0.99	0.86	1.21	0.72	0.000	0.64
university level (UK)		0.80	0.47	0.92	0.54	0.000	0.84
deeper school knowledge (DSK)		0.69	0.47	0.96	0.31	0.000	0.99

Table 2 Development of PCK and PK (sum-scores, in %)

	N	t_1		t_2			
PCK	63	M	SD	M	SD	p	d
experiments		0.50	0.20	0.55	0.23	0.106	0.21
physics education concepts		0.54	0.18	0.63	0.14	0.000	0.59
instructional strategies		0.38	0.19	0.39	0.16	0.740	0.04
students' misconceptions		0.48	0.18	0.53	0.20	0.043	0.26
PK	58	M	SD	M	SD	p	d
		0.35	0.13	0.44	0.11	0.000	0.82

Regarding CK, a significant increase in all subdimensions can be found (medium to large effect, *Cohen's d* ranging from 0.64 to 0.99, Table 1). Regarding PCK, significant increases can be seen in the sub-scales "student's misconceptions" (small effect, *Cohen's d* = 0.26, Table 2) and "physics education concepts" (medium effect, *Cohen's d* = 0.59, Table 2). Furthermore, a significant increase with a large effect occurred regarding PK (*Cohen's d* = 0.82, Table 2).

Table 3 contains the results of the performance assessments. All scores represent manifest scores (Sections 5.4 to 5.6). As Table 3 shows, only for one out of the three performance assessments a significant increase with a small effect can be observed so far (planning physics lessons, *Cohen's d* = 0.32).

Table 3 Development of skills regarding three standard teaching situations (sum-scores, in %)

	N	t ₁		t ₂		p	d
		M	SD	M	SD		
planning physics lessons	66	0.49	0.11	0.53	0.13	0.01	0.32
reflecting on physics teaching	37	0.11	0.03	0.12	0.04	0.404	0.16
explaining physics	20	0.30	0.15	0.33	0.12	0.420	0.24

7 Discussion

The newly developed performance assessments enable valid interpretations (Vogelsang et al. 2019) of the data reflecting three important skills of prospective physics teachers: planning and reflecting on physics lessons, and explaining physics. Regarding the one-semester internship, significant increases in all sub-dimensions of CK, in two sub-dimensions of PCK, and for PK could be observed. Due to fewer (formal) learning opportunities during the whole internship program (teacher training at schools and accompanying courses at a university) we expected little increase in more theoretical professional knowledge. However, we found medium to large effects in terms of professional knowledge. In addition, contrary to our expectations, no increases of the skills to reflect on physics teaching and to explain physics could be identified so far. That leads to the question of whether typical long-term school internships programs really contribute to acquiring professional skills, especially regarding the expected application of theoretical knowledge for high-quality performance in teaching situations (which is often expected, cf. Cohen et al. 2013). The non-significant development of skills for reflecting and

explaining measured by standardized assessments may indicate a lack of effectiveness of the internship regarding central goals.

It should be noted that the results presented are preliminary and based on small sample sizes – the statistical power seems to be limited. The collected data have not been coded completely up til now and some of the coding categories of the performance assessments for planning and reflecting will be finetuned for further analyses. Although the sample is small—due to the small number of physics teacher students in Germany in general—it should be noted that we were able to carry out almost complete surveys at the participating four universities.

Other limitations lie in the low reliabilities of some sub-scales of the knowledge assessments and in the score regarding the skills to reflect on physics teaching. Besides, further improvements of the codings for the performance assessments have to be done, which might lead to an increase of the reliability and might help to identify smaller effects. For all instruments, core arguments for the construct validity will also be checked based on an extensive nomological network. Furthermore, the use of performance assessments following approaches from medical education is quite uncommon in teacher education at least in Germany (see also Kuhn et al. in this volume). One possible reason, why no effects could be observed, might lie in this uncommon testing situations for the participating students. Another reason might be a decrease in the test motivation over the internship due to the long-lasting data collection. However, such a decrease could only be identified between the measurement points of the assessment for reflection skills (based on the results of a short test-motivation scale, Vogelsang et al. 2019).

Further analyses are needed to shed more light on the effectiveness of long-term internships in teacher education. We tried to research changes in professional knowledge and skills before and after the intership using pre-post-measurements. Future studies should take a closer look at the learning processes of the student teachers *during* the internship. Although we have information regarding the content of the accompanying courses at the universities and, additionally, self-reports of the prospective teachers on their experiences with their mentors during the intership, there is a lack of information on learning processes in detail. Future studies focussing on changes in shorter time periods linked to training situations could be a useful additional approach.

Funding

Profile-P+ was funded by the KoKoHs program of the German Ministry for Education and Research (01PK15005A-D).

References

- Besa, K. S., & Büdcher, M. (2014). Empirical evidence on field experiences in teacher education. In K. H. Arnold, A. Gröschner, & T. Hascher (Eds.), (2014), *Schulpraktika in der Lehrerbildung: Theoretische Grundlagen, Konzeptionen, Prozesse und Effekte* (pp. 129–145). Waxmann.
- Blömeke, S., Gustafsson, J. E. & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223, 3–13.
- Carlson, J., Daehler, K.R., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., & Wilson, C. (2018). The Refined Consensus Model of pedagogical content knowledge in science education. Repositioning pedagogical content knowledge in teachers' knowledge for teaching science. In A. Hume, R. Cooper & A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77–92). Singapore: Springer.
- Cauet, E., Liepertz, S., Kirschner, S., Borowski, A., & Fischer, H. E. (2015). Does it matter what we measure? Domain-specific professional knowledge of physics teachers. *Revue Suisse des sciences de l'éducation*, 37(3), 462–479.
- Clarà, M. (2015). What Is Reflection? Looking for Clarity in an Ambiguous Notion. *Journal of Teacher Education*, 66(3), 261–271.
- Cohen, E., Hoz, R., & Kaplan, H. (2013). The practicum in preservice teacher education: a review of empirical studies. *Teacher Education*, 24, 345–380.
- Cortina, K. S., & Thames, M. H. (2013). Teacher education in Germany. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 49–62). Boston: Springer.
- Diez, M. E. (2010). It Is Complicated: Unpacking the Flow of Teacher Education's Impact on Student Learning. *Journal of Teacher Education*, 61(5), 441–450.
- Geelan, D. (2012). Teacher explanations. In B. Fraser, K. Tobin, & C. McRobbie (Eds.), *Second international handbook of science education* (pp. 987–999). Dordrecht: Springer.
- Gramzow, Y., Riese, J., & Reinhold, P. (2013). Prospective physics teachers' pedagogical content knowledge – Validating a test instrument by using a think aloud study. *European Science Education Research Association 2013 Conference*, Nicosia, Cyprus.
- Harden, R., Stevenson, M., & Wilson, W. G. W. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1, 447 – 451.
- Holtz, P., & Gnamb, T. (2017). The improvement of student teachers' instructional quality during a 15-week field experience: a latent multimethod change analysis. *Higher Education*, 74(4), 669–685.
- Kirschner, S. (2013). *Modellierung und Analyse des Professionswissens von Physiklehrkräften [Modeling and Analyzing Professional Knowledge of Pre-Service Physics Teachers]*. Berlin: Logos.
- König, J., Blömeke, S., Paine, L., Schmidt, W. H., & Hsieh, F. J. (2011). General pedagogical knowledge of future middle school teachers: On the complex ecology of teacher education in the United States, Germany, and Taiwan. *Journal of Teacher education*, 62(2), 188–201.
- Korneck, F., Krüger, M. & Szogs, M. (2017). Professionswissen, Lehrertüberzeugungen und Unterrichtsqualität angehender Physiklehrkräfte unterschiedlicher Schulformen [Profes-

- sional knowledge, teacher beliefs and instructional quality of prospective physics teachers of different school types]. In H. Fischler & E. Sumfleth (Eds.), *Professionelle Kompetenz von Lehrkräften der Chemie und Physik [Professional Competence of chemistry and physics teachers]* (pp. 113–133). Berlin: Logos.
- Kulgemeyer, C., & Riese, J. (2018). From professional knowledge to professional performance: The impact of CK and PCK on teaching quality in explaining situations. *Journal of Research in Science Teaching*, *55*(10), 1393–1418.
- Kulgemeyer, C., & Tomczyszyn, E. (2015). Physik erklären – Messung der Erklärens-fähigkeit angehender Physiklehrkräfte in einer simulierten Unterrichtssituation [Explaining physics – Measuring teacher trainees’ explaining skills using a simulated teaching setting]. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 111–126.
- Lee, E., & Luft, J. A. (2008). Experienced secondary science teachers’ representation of pedagogical content knowledge. *International Journal of Science Education*, *30*(10), 1343–1363.
- Liepert, S., & Borowski, A. (2018). Testing the Consensus Model: relationships among physics teachers’ professional knowledge, interconnectedness of content structure and student achievement. *International Journal of Science Education*. doi: 10.1080/09500693.2018.1478165
- Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95 – 132). Dordrecht: Kluwer Academic Publishers.
- Mayring, P. (2000). Qualitative Content Analysis. *Forum: Qualitative Social Research*, *1*(2), 1–10.
- Mertens, S., & Gräsel, C. (2018). Entwicklungsbereiche bildungswissenschaftlicher Kompetenzen von Lehramtsstudierenden im Praxissemester [The development of educational competences during long-term internships in teacher education]. *Zeitschrift für Erziehungswissenschaft*, *21*(6), 1109–1133.
- Miller, G.E. (1990). The Assessment of Clinical Skills/Competence/Performance. *Academic Medicine*, *64*(9), 63–67.
- Nowak, A., Kempin, M., Kulgemeyer, C., & Borowski, A. (2019). Reflexion von Physikunterricht [Reflection of Physics Teaching]. In C. Maurer (Eds.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe [Scientific Education as a basis for vocational and social participation]* (pp. 838–841). University of Regensburg.
- Ohle, A., Boone, W. J., & Fischer, H. E. (2014). Investigating the impact of Teachers’ Physics CK on Students outcomes. *International Journal of Science and Mathematics Education*, *13*(6), 1211–1233.
- Plöger, W., Scholl, D., & Seifert, A. (2015). Analysekompetenz – ein zweidimensionales Konstrukt?! Unterrichtswissenschaft [Analytical competence – a two-dimensional construct?!]. *Zeitschrift Für Lernforschung*, *43*(2), 166–184.
- Riese, J., & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen [Professional competence of student teachers enrolled in different teacher training programs]. *Zeitschrift für Erziehungswissenschaft*, *15*(1), 111–143.

- Riese, J., Kulgemeyer, C., Zander, S., Borowski, A., Fischer, H., Gramzow, Y., Reinhold, P., Schecker, H. & Tomczyszyn, E. (2015). Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik [Modelling and measurement of professional knowledge in academic physics teacher education]. In S. Blömeke, & O. Zlatkin-Troitschanskaia (Eds.), *Kompetenzen von Studierenden: 61. Beiheft der Zeitschrift für Pädagogik [Competences of students in higher education]* (pp. 55–79). Weinheim: Beltz.
- Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The influence of teachers' knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, 50(5), 1020 – 1049.
- Seifert, A., & Schaper, N. (2012). Die Entwicklung von bildungswissenschaftlichem Wissen: Theoretischer Rahmen, Testinstrument, Skalierung und Ergebnisse [The development of Pedagogical Knowledge: Theoretical Framework, Test Instrument, Scaling and Results]. In J. König & A. Seifert (Eds.), *Lehramtsstudierende erwerben pädagogisches Professionswissen – Ergebnisse der Längsschnittstudie LEK [Pre-service teachers acquire pedagogical professional knowledge – results of the longitudinal study LEK]* (pp. 183–214). Münster: Waxmann.
- Shavelson, R. J., & Stern, P. (1981). Research on Teachers' Pedagogical Thoughts, Judgments, Decisions, and Behavior. *Review of Educational Research*, 51(4), 455–498.
- Shulman, L. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4–14.
- Stender, A. (2014). *Unterrichtsplanung: Vom Wissen zum Handeln. Theoretische Entwicklung und empirische Überprüfung des Transformationsmodells der Unterrichtsplanung [Lesson planning: From knowledge to action. Theoretical development and empirical verification of the transformation model of lesson planning]*. Berlin: Logos.
- van Driel, J., Biejaard, D. & Verloop, N. (2001). Professional Development and Reform in Science Education – The Role of Teachers' Practical Knowledge. *Journal of Research in Science Teaching*, 38(2), 137–158.
- Vogelsang, C., Bowoski, A., Buschhüter, D., Enkrott, P., Kempin, M., Kulgemeyer, C., Reinhold, P., Riese, J., Schecker, H. & Schröder, J. (2019). Entwicklung von Professionswissen und Unterrichtsperformanz im Lehramtsstudium Physik – Analysen zu valider Testwertinterpretation [Development of Professional Knowledge and teaching skills in Academic Physics Teacher Education – analyses regarding the valid interpretation of test scores]. *Zeitschrift für Pädagogik*, 65(4), 473–491.
- Volmer, M., Pawelzik, J., Todorova, M. & Windt, A. (2019). Reflexionskompetenz von Sachunterrichtsstudierenden im Praxissemester. In C. Maurer (Ed.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe [Scientific Education as a basis for vocational and social participation]* (pp. 321–324). University of Regensburg.
- Woitkowski, D. & Borowski, A. (2017). Fachwissen im Lehramtsstudium Physik [Content Knowledge in Academic Physics Teacher Education]. In H. Fischler & E. Sumfleth (Eds.), *Professionelle Kompetenz von Lehrkräften der Chemie und Physik [Professional Competence of chemistry and physics teachers]* (pp. 57–76). Berlin: Logos.



2.6 Linguistically Responsive Teaching in Multilingual Classrooms

Development of a Performance-Oriented Test to Assess Teachers' Competence

Lemmrich, S., Hecker, S.-L., Klein, S., Ehmke, T., Koch-Priewe, B., Köker, A., and Ohm, U.

Abstract

The *DaZKom-Video* project aims at performance-oriented measurement of German-as-a second-language (GSL)-competence of pre- and in-service teachers using video-based stimuli and oral responses. The present study is part of this project and focuses on the psychometric quality of the *DaZKom-Video* test instrument and evaluates the dimensionality of the construct *GSL competence (Deutsch-als-Zweitsprache-Kompetenz)*. Additionally, we explore correlations between GSL competence and sociodemographic data as well as learning opportunities. The sample consists of pre- and in-service teachers from different school forms and universities across Germany. The items and scale analyses reveal good psychometric quality: The item-fit to the Rasch-model as well as the discrimination is satisfactory. The dimensional analyses show best model fit for the one-dimensional and three-dimensional models. Test persons with GSL experience such as in-service teachers (in contrast to pre-service teachers), test persons with teaching experience, those with an additional qualification in the field of GSL and test persons with many learning opportunities are statistically more likely to respond to items correctly. These results indicate a reliable and valid test instrument for performance-oriented measurement of GSL competence.

Keywords

Video-based testing, teacher education, linguistically responsive teaching, teachers' competence

1 Introduction

Language is crucial in teaching and learning: Every subject-specific classroom uses language to represent content. Language has specific characteristics in content classrooms and every subject adds its own subject-specific register (e.g. written language or specific operators, Schmölzer-Eibinger 2013). For some pupils, especially but not only for those with German as a second language (GSL), it therefore can be challenging to apprehend content successfully. Teachers of all subjects must be aware that specific registers come with their own demands and may pose challenges for their pupils. Accordingly, teachers need competencies to support their pupils in gaining access to the content by teaching linguistically responsive (Thürmann and Vollmer 2013, Cummins 2001). Kimanen et al. (2019) implicate that teachers' competence in linguistically and culturally responsive teaching could increase teachers' self-esteem and therefore have an impact on teachers' motivation and commitment to teaching. How teachers must be trained concerning teaching multilingual learners is still unclear, however, and performance standards of teachers' competencies in this field are missing. Furthermore, understanding which learning opportunities may contribute to the development of teachers' competencies in teaching in linguistically diverse classrooms is still a desideratum (Baumann 2017).

The *DaZKom-Video*¹ project aims at evaluating and improving learning opportunities for pre- and in-service teachers in teaching multilingual learners and developed a performance-oriented test instrument. The theoretical and assessment framework includes the GSL competence model and the paper-pencil-test of the previous project (*DaZKom*²). To measure teachers' professional competencies and not solely capture teachers' knowledge, but operate as closely to performance as possible, it is necessary to use performance-oriented test formats (Aufschnaiter

- 1 The project *DaZKom-Video: Performanznahe Messung von Deutsch-als-Zweit-sprache-Kompetenz bei (angehenden) Lehrkräften (2017–2020)*.
- 2 The previous project *DaZKom (2012–2015)* was funded by the German Federal Ministry for Education and Research (BMBF) (01PK11010A, 01PK11010B).

and Blömeke 2010). Consequently, the test instrument uses video vignettes of teaching situations and asks for orally given responses that are then recorded.

This article examines the extent to which the developed performance-oriented test instrument is suitable for capturing the construct of GSL competence. It evaluates different scaling models concerning the dimensionality of the GSL construct. As indicators for convergent and discriminant validity, the correlations of the measured GSL competence and socio-demographic characteristics as well as GSL-related learning opportunities of test persons are examined in this study.

Concerning the terminology, this paper uses ‘teaching pupils with German as a second language (GSL)’ synonymously to ‘teaching multilingual learners’ and ‘teaching in linguistically diverse classrooms’. GSL competence therefore refers to the competency to teach multilingual learners/in linguistically diverse classrooms.

2 Theoretical Background

2.1 Structure of GSL Competence

A structural model of teachers’ GSL-competence served as the theoretical framework and basis for the performance-oriented test development. This model was devised in a previous project on the basis of document analysis, followed by an expert rating (Ehmke and Hammer 2018; Köker et al. 2015). Due to the premise that GSL competency is relevant for teaching all subjects in the mainstream classroom, the model consists of different partial competencies which are important for subject content integrated language facilitation and dealing adequately with multilingualism. The model illustrates the question of how subject content can be taught in such a way that the linguistic needs of all pupils, especially those with German as their second language, are taken into account (Carlson et al. 2018). In addition, the model depicts what teachers need to know to be able to act linguistically responsive. This knowledge refers to linguistic aspects, for instance, with regard to the specificity of specialized texts, aspects of second language acquisition and multilingualism as well as the didactic aspects of lesson planning and implementation. Since GSL competence is considered a generic competence in all subjects, it always refers to subject content.

The structural model discerns three dimensions: *subject-specific register*, *multilingualism*, and *didactics* (Ohm 2018). All dimensions are differentiated into subdimensions and further characterized by content-related facets.

The first dimension, *subject-specific register*, concerns language as a medium for classroom interaction and understanding technical concepts as well as a

learning object. It is assumed that all subjects have their specific registers that are essential for the students' knowledge construction and must therefore be taught explicitly in connection with subject contents (Schleppegrell 2004).

The second dimension, *Multilingualism*, includes the two subdimensions *Second Language Acquisition* and *Migration* and focuses on the learning process (Ohm 2018). To support the (GSL) students according to their individual language learning levels, teachers require knowledge about the development of students' linguistic competence, for instance, regarding milestones of second language acquisition. Second language learning resp. the increasingly differentiated use of linguistic registers occur in the context of multilingualism. At the same time, multilingualism refers not only to external but also to internal multilingualism, such as the use of dialects. Therefore, the second dimension considers both linguistic diversity in school and dealing with heterogeneity.

The third dimension, *Didactics*, includes the subdimensions *formative assessment* in classrooms and *language facilitation* (Ohm 2018) and complements the first two dimensions. Based on knowledge about subject-specific registers and the specific characteristics of second language development, teachers plan and conduct their lessons. Consequently, the teaching process is at the focus of this dimension.

2.2 Indicators of Performance: Perception and Decision-Making

The situation-specific skills, *perception* and *decision-making* (Blömeke et al. 2015), serve as indicators of performance in GSL-relevant teaching situations. In research on teacher professionalization, perception often is defined by referring to the concept of *professional vision* by Goodwin (1994). He explains experts perceiving meaningful situations in their domain of expertise more selectively than novices do, as experts' perception has been trained to that effect in years of practice. Van Es and Sherin (2002) define two components of teachers' *professional vision*: *noticing* and *knowledge-based reasoning*. While the first term describes the filtered selective perception, the second refers to the ability to draw conclusions based on the situations perceived (Es and Sherin 2002). Therefore, experts are able to resort to their professional knowledge to classify the perceived into concepts and theories and react more precisely. In contrast, novices often deliver only a mere description of what they observed (Seidel et al. 2010).

Professional competence consists of professional knowledge and individual abilities, as well as motivational, volitional and social aspects and is dependent on domain-specific requests of action (Weinert 2001; Baumert and Kunter 2006).

Lindmeier (2013) captures the ability of *decision making* as *action-oriented* competency and thereby refers to the ability of teachers to use their professional knowledge spontaneously and immediately in situations (ibid.; see also Kuhn et al. in this volume).

The *DaZKom-Video* test instrument uses test items which pick up this theoretical structure. Video-stimuli present authentic teaching situations with specific *perception* and *decision-making* tasks to measure the German-as-a-Second-Language competence of the (pre-service) teachers as closely to performance as possible. The video-vignettes were carefully chosen after reviewing many teaching situations (video and audio recordings) and were validated for their authenticity by experts of universities and teacher experts in schools.

3 Research Questions

Based on the theoretical framework, three research questions will be answered in this paper:

1. How satisfactory is the *DaZKom-Video* test instrument's psychometrical quality?

Based on the evaluation of data of a first pilot study (Lemmrich et al. 2019), it can be assumed that the data will have a good item fit to the Rasch-model. Additionally, the discrimination in the first pilot study was satisfactory, even though the items were too difficult for the respective sample then.

2. Which dimensional structure can be determined for the test instrument?

Considering the results of the paper-pencil test of the previous project (*DaZKom*) it can be assumed that the construct *GSL competence* is structured multidimensionally (Hammer et al. 2015), also in a performance-oriented test. However, taking the results of the first pilot study into account (Lemmrich et al. 2019), it is possible that the theoretically determined dimensions do not appear clearly separated. As the two-dimensional analysis (*perception (1)* and *decision-making (2)*) showed strong correlations between these two dimensions in the first pilot study (ibid.) we expect a similar result in the second pilot study. Referring to *GSL competence*, three dimensions can be determined based on the theoretical assumptions (*subject-specific register (1)*, *multilingualism (2)*, *didactics (3)*, see *DaZKom-Model* in Köker et al. 2015, Ohm 2018; Section 2), which either appear as one-dimensional

or can be determined separately. In the first pilot study, results showed a medium correlation between dimension 1 and 3 and a low correlation between both of these and dimension 2. As the present pilot study was carried out with a revised test instrument and a different sample including more in-service teachers than the first pilot study, the results are expected to clarify the dimensionality of the construct.

3. Which correlations can be identified between the GSL competence of the test persons, the sociodemographic characteristics and learning opportunities?

The correlations examined give insight into the convergent and discriminant validity. We assume that test persons who had GSL-relevant learning opportunities are more likely to respond correctly. Additionally, experienced teachers are expected to pass the performance-oriented test more successfully than unexperienced teachers or pre-service teachers.

4 Methods

4.1 Test Instrument

The project team collected data in a standardized way at four universities and six schools in four German federal states (Berlin, Lower Saxony, North-Rhine Westphalia, Hamburg). The conductors first introduced the test closely following a manual, covering both the test procedure as well as the test's background. In their introductions, the conductors invariably pointed out the items' GSL focus. Afterwards, the test persons took two of four sub tests. Each subtest contained four video stimuli. Headsets and tablets were provided to enable the participants to take the performance-oriented test independently. After watching an example item including a sample solution, eight video-vignettes were shown, each followed by two items on the situation-specific skills *perception* and *decision-making*. The video-vignettes used as prompts take 30 seconds to three minutes, each dealing with authentic, GSL-relevant teaching situations that were taken from different subject lessons at secondary school classes. Subjects are Mathematics, German, Ethics and Science. All situations were matched both to the dimensions and subdimensions of the DaZKom-Model (Köker et al. 2015). Their adequacy, GSL relevance and typicality for the chosen dimension were assessed and approved in expert ratings with $N = 3$ university experts in the field.

To best simulate the immediacy and spontaneity of teacher behavior in real situations and thereby achieve a higher ecological validity, only open-response

items on the test persons' perception (*What do you perceive?*) and on their simulated (re-)action (a. *You are the teacher in this situation, how do you react word for word?*, or b. *How would you act in this situation if you were the teacher?*) were used, each asking for an oral response using the headsets' built-in microphones. To enable these kinds of open responses, the video vignettes usually closed with a student's statement the test persons could react to. At last, the test persons took an additional questionnaire both on their learning opportunities and sociodemographic data (Section 4.3).

4.2 Coding Manual

The coding manual, required to measure difference in quality in the test persons' responses reliably, was developed in a complex multi-step process. It included expert ratings ($N = 6$) with $n = 3$ researchers from the field (PhD and above) and $n = 3$ experienced teachers (different subjects, 30+ years of experience in schools while also working as instructors offering further training in GSL). The experts were taken through the same procedure as the regular test persons: they first watched the video vignettes before giving oral responses to the items on their situation-specific skills, independently and without a project member being present. Their responses were audio-taped and afterwards used for a qualitative content analysis employed to identify codes (Nimz, Hecker & Köker 2018). These codes were then ascribed to the scores 0, 1 and 2 in the project team. Based on this allocation, the actual item specific coding manuals were designed. Several rounds of optimization followed. The manuals were tested using data from the first pilot study conducted in Germany ($N = 137$; $n = 40$ in-service teachers, $n = 83$ pre-service teachers, $n = 11$ scientists). All responses were double coded by two independent raters. During the rating, doubtful cases were identified and discussed in the project team, more anchors added and necessary adjustments to the manuals made. Finally, the team achieved a satisfactory interrater agreement (Cohen's Kappa $\kappa = .76$). The final manuals include closed descriptions of the values in addition to many anchors. In the process of scoring the responses, only the best aspects in every response count, the rest is not taken into account. A maximum of two points can be reached in each item. For the analysis performed in this study, code 2 and code 1 were merged into code 1.

4.3 Additional Questionnaires: Sociodemographic Data and Learning Opportunities

After giving their oral responses, the test persons filled in additional questionnaires on sociodemographic characteristics and learning opportunities. The first questionnaire presented questions on gender, subjects of studies, teaching experience, etc. (6 items). The learning opportunities questionnaire contains two scales: One scale of 16 items on GSL-relevant topics that the test persons might have discussed in their studies/teacher training ($\alpha = 0.91$; Ehmke and Lemmrich 2018). Furthermore, it contains a scale of eight items ($\alpha = 0.83$, *ibid.*) on GSL-relevant actions the test persons might have taken before in their learning contexts (e.g. studies, further teacher training, etc.). Both scales use a five-point Likert scale ((1) *never* (2) *in one session*, (3) *in several sessions*, (4) *in a module*, (5) *in several modules*). Additionally, there were three single items that aimed for answers on GSL-specific teaching experience, additional qualifications and experience in research in the field of teaching multilingual learners.

4.4 Sample

The sample size of this study was $N = 184$ test persons, 79% female and 21% male; 51.2% were pre-service teachers (teacher students) from German universities and 39.6% were in-service teachers from different areas of Germany and different types of schools; 8.2% were “other”, such as scientists, teacher educators and students of other disciplines. The test persons studied/taught different kinds of subjects (Math, German, English, Physics, Biology, Chemistry, Geography, Social Studies/Science, History, Music, Arts, Religion, Politics, Sports); 25.6% had additional qualifications in the field of GSL/teaching multilingual learners (e.g. GSL-certificate); 88.6% had German as their first language, while 11.4% of the test persons had a different first language, which mostly were Turkish (3.8%) and Russian (1.6%).

4.5 Statistical Procedure

The collected data (coding of the *DaZKom-Video* test, sociodemographic data and learning opportunities) were analyzed with SPSS 25 (IBM 2018), including frequencies and information on the sample. The psychometric analysis of items and scales was carried out in *ConQuest* (Adams et al. 2015) on the basis of the

Rasch-model (Rost 2004). The GSL competence of the test persons was determined with the Weighted Likelihood Estimates (WLEs). WLEs take two factors into account: Firstly, the correct reply to an item depending on the respondents' abilities (in this case GSL competence); secondly, the item difficulty (Wilson 2005). Hence, the WLE does not map the actual responses of the test persons, but the probability of a correct answer based on the test persons' skills and the item difficulty (ibid.). To examine the dimensionality of the construct GSL-competence, a one, a two and a three-dimensional model were estimated with *ConQuest*.

5 Results

5.1 Psychometric Quality

Table 1 shows the results of the IRT-Scaling. The first criterion for psychometric quality is the item fit based on the Rasch-model, which is very good ($1.0 < MNSQ^3 < 1.25$; OECD 2005). A few items show a lower item fit ($Min = 0.88$; $Max = 1.10$), but at least the maximum item fit is located within the satisfactory range. The second criterion is discrimination, which indicates how reliable the instrument distinguishes between test persons with high and low ability estimates. Three items that showed a discrimination of under 0.3 were excluded from all analysis. Thereafter, the average discrimination of all items was satisfactory ($M = 0.43$; $SD = 0.09$). The overlap of the average item difficulty ($M = 0.88$; $SD = 0.78$) and the respondents' abilities ($M = 0.00$; $SD = 1.08$) is not fully satisfactory yet. Therefore, the items either are too difficult or the respondents' abilities not sufficient enough. The EAP-reliability of the one-dimensional scaling amounts $\alpha = 0.62$.

3 *Weighted Meansquare Fit Statistics (MNSQ)*: indicates if the expected and the observed likelihood to respond correctly relate.

Table 1 Item and scale indices

	M	SD	Min	Max
Weighted fit MNSQ	1.00	0.05	0.88	1.10
Discrimination	0.43	0.09	0.30	0.58
Item difficulty	0.88	0.78	-0.53	2.79
WLE	0.00	1.08	-2.87	3.28

5.2 Verification of Dimensions

The dimensionality of the test instrument was tested with one-, two- and three-dimensional IRT models. Table 2 shows the fit-indices of the comparison of the three models: the deviance, the AIC, BIC and the CAIC. All four of them are criteria that give information about the goodness of fit of the estimated model. They take into account the empirical data as well as the model (Moosbrugger and Kelaya 2007, p. 390). All these criteria should be as low as possible (*ibid.*). The BIC and the CAIC are at the lowest level for the one-dimensional model. The deviance and the AIC are at lowest for the three-dimensional model. The construct *GSL competence* therefore is likely to be a one-dimensional or three-dimensional construct rather than a two-dimensional construct. The reliability of the three-dimensional model was not satisfactory due to a very low number of items per dimension. The following analyses therefore are operated with data of the one-dimensional model.

Table 2 Comparison of dimensions

	N	Parameter	Deviance	AIC	BIC	CAIC
1-dimensional	176	30	2606	2666	2761	2791
2-dimensional	176	32	2607	2671	2773	2805
3-dimensional	176	35	2585	2655	2766	2801

Looking at the two-dimensionally modeled construct, the two assumed dimensions *perception* and *decision-making* (Section 2.2) highly correlate ($r = 0.97$, Figure 1). It therefore can be assumed that the construct *GSL competence* is not separable into these two dimensions. The three-dimensional modeled construct (Section 2.1) shows moderate correlations: dimension 1 (*subject specific register*) and dimension 2 (*multilingualism*) correlate by $r = 0.46$, dimension 2 and 3 (*didactics*) by $r = 0.51$ and dimension 1 and 3 by $r = 0.55$. This result indicates that these three theoretical dimensions could be separated within the construct *GSL competence*.

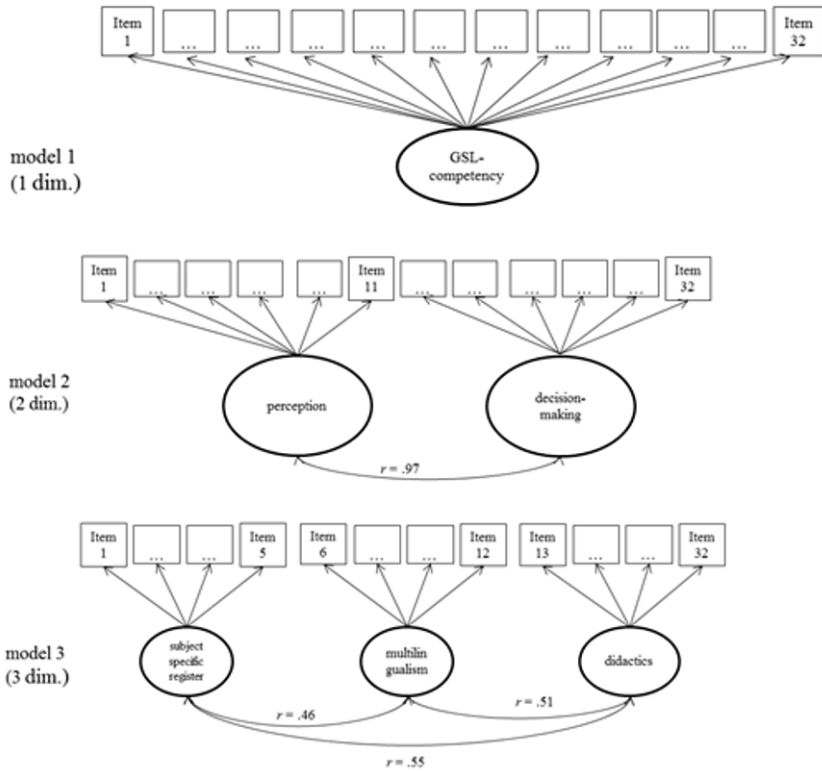


Figure 1 Comparison of models

5.3 Correlations between GSL Competence, Personal Characteristics and Learning Opportunities

Table 3 shows correlations between sociodemographic characteristics and GSL-relevant learning opportunities with the weighted likelihood estimates (WLE). There is no statistically significant difference between the female and male test persons probability to respond correctly. The results show a tendency that test persons with German as their first language are more likely to respond correctly, even though the effect is statistically non-significant. Teachers' responses are more often correct than those of students' (pre-service teachers). This effect is confirmed by the statistically significant correlation of teaching experience with the

WLE. Also, test persons who have a GSL-certificate, which is a specific qualification for teaching multilingual learners, are more likely to respond correctly than those without. Lastly, learning opportunities (sum score of both scales) correlate with the WLE: The more GSL-relevant learning opportunities test persons had, the more likely they were to respond correctly.

Table 3 Correlations between the results of the DaZKom-Video-test and sociodemographic data

	WLE
Gender (male = 0; female = 1)	0.14
Status (teacher student = 0; teacher = 1)	0.04
GSL – certificate ¹⁾ (no = 0, yes = 1)	0.20**
Teaching experience GSL ²⁾ (no = 0; yes = 1)	0.16*
Research GSL ³⁾ (no = 0; yes = 1)	0.09
Scale GSL-learning opportunities ⁴⁾	0.18*

* The correlations are statistically significant at a level of .05 (two-sided).

** The correlations are statistically significant at a level of .01 (two-sided).

- 1) additional certificate that proves the attendance of a specific course on teaching students with German as a second language
- 2) experience in teaching students with German as a second language
- 3) research experience in the field of German as a second language/teaching multilingual learners
- 4) learning opportunities in the field of German as a second language/teaching multilingual learners

6 Discussion and Outlook

This study has been carried out to evaluate the psychometric quality and the dimensionality of the performance-oriented test instrument. Additionally, correlations of GSL competence with sociodemographic data and learning opportunities were examined. The main results were the following:

1. The test instrument shows a good fit to the Rasch-model. Three items were excluded from the test due to a low discrimination value (< 0.3). The discrimination was satisfactory. At the same time, the test did not fit perfectly to the ability of the test persons, which indicates that the items were too difficult.
2. The dimensional analyses showed the best model fits for the one-dimensional and the three-dimensional models. The three-dimensional structure represents

the theoretical structure of the GSL assessment framework. The two dimensions of perception and decision making could not be separated statistically from each other.

3. The correlations between the WLE of the test persons concerning GSL competence and the sociodemographic data/learning opportunities indicate a higher likelihood of responding correctly for test persons who are experienced teachers. Teaching experience as well as more learning opportunities correlated statistically significant with the WLE.

The psychometric analysis showed that some items were still too difficult for the test persons. However, compared to the first pilot study (Lemmrich et al. 2019), the item fit increased. After the first evaluation, the test instrument was revised: Stimuli that did not work were excluded and more context information was added to increase the specificity of the situations. Additionally, the current sample consists of half students (pre-service teachers) and half teachers, whereas the first pilot study was carried out with a majority of students, who are less experienced than teachers and therefore not as likely to perform successfully in a performance-oriented test instrument (Section 2).

Concerning the dimensionality of the construct, there is no evidence for a two-dimensional construct. *Perception* and *decision-making* cannot be determined separately for the construct. This result is in line with results of a first pilot study (Lemmrich et al. 2019) and is compliant with theories of cognition psychology as well as other studies on professional vision and performance (Niesen 2018; Hommel et al. 2019). Another possible explanation for the results is the fact that the items corresponding to either *perception* or *decision-making* both demand answers that are based on one and the same video vignette. The results indicate a one-dimensional construct in a performance-oriented environment. However, in the three-dimensional modeling (three dimensions of GSL competence: *subject specific register*, *multilingualism* and *didactics*; Section 2), the three dimensions show only medium correlations, which indicates that a differentiated analysis might be reasonable. Further analysis and evaluation will examine the dimensionality in detail with the final test instrument and a larger sample with a higher number of teachers. The standardization study will examine a sample of $N = 300$ ($N_{\text{teachers}} = 150$) with a revised test instrument (12 final video vignettes in one test version).

Looking at the correlations of GSL-competence with sociodemographic data and learning opportunities, the results comply with the theoretical background. Experience and learning opportunities correlate with the WLEs of the test persons: test persons with teaching experience as well as with additional qualifica-

tions in the field of GSL or learning opportunities are more likely to respond correctly to the items.

The instrument contains 16 vignettes that show very short teaching examples. It therefore offers only a selection of possible contexts and cannot portray the broad variety and complexity of everyday teaching situations. This study aimed at evaluating the psychometric quality and dimensionality of the test instrument and did not analyze descriptive values. Future studies (with an equally high psychometric quality) will take this into account and identify GSL-relevant learning opportunities to then comprehensively train teachers in this field and prepare them optimally to teach in linguistically diverse classrooms.

Funding

DaZKom-Video was funded by the KoKoHs-program of the German Federal Ministry for Education and Research (01PK16001A, 01PK16001B).

References

- Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised Item Response Modelling Software [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.
- Aufschnaiter, C. v., & Blömeke, S. (2010). Professionelle Kompetenz von (angehenden) Lehrkräften erfassen – Desiderata. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, pp. 361–367.
- Baumann, B. (2017). Sprachförderung und Deutsch als Zweitsprache in der Lehrbildung. In M. Becker-Mrotzek, P. Rosenberg, C. Schroeder & A. Witte (Eds.), *Deutsch als Zweitsprache in der Lehrerbildung* (pp. 9–26). Münster: Waxmann.
- Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9* (4), (pp. 469–520).
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, *223* (1), pp. 3–13.
- Carlson, S. A., Köker, A., Rosenbrock-Agyei, S., Ohm, U., Koch-Priewe, B., Hammer, S., Fischer, N., & Ehmke, T. (2018). DaZKom – a Structure Model of Pre-service Teachers' Competency for Teaching German as a Second Language in the Mainstream Classroom. In T. Ehmke, S. Hammer, A. Köker, U. Ohm & B. Koch-Priewe (Eds.), *Professionelle Kompetenzen angehender Lehrkräfte im Bereich Deutsch als Zweitsprache* (pp. 261–283). Münster and New York: Waxmann.
- Cummins, J. (2001). *Negotiating Identities: Education for Empowerment in a Diverse Society*. Second Edition. Los Angeles: California Association for Bilingual Education.

- Ehmke, T., & Hammer, S. (2018). Skalierung und dimensionale Struktur des DaZ-Kom-Testinstruments. In T. Ehmke, S. Hammer, A. Köker, U. Ohm & B. Koch-Priewe (Eds.), *Professionelle Kompetenzen angehender Lehrkräfte im Bereich Deutsch als Zweitsprache* (pp. 129–148). Münster: Waxmann.
- Ehmke, T., & Lemmrich, S. (2018). Bedeutung von Lerngelegenheiten für den Erwerb von DaZ-Kompetenz. In T. Ehmke, S. Hammer, A. Köker, U. Ohm & B. Koch-Priewe (Eds.), *Professionelle Kompetenzen angehender Lehrkräfte im Bereich Deutsch als Zweitsprache* (pp. 201–219). Münster: Waxmann
- Goodwin, C. (1994). Professional vision. *American Anthropologist*, 96 (3), pp. 606–633.
- Hammer S., Carlson, S. A., Ehmke, T., Koch-Priewe, B., Köker, A., Ohm, U. Rosenbrock, S., & Schulze, N. (2015). Kompetenz von Lehramtsstudierenden in Deutsch als Zweitsprache. Validierung des GSL-Testinstruments. *Zeitschrift für Pädagogik*, Beiheft 61, pp. 32–54.
- Hommel, B., Müsseler, J., Aschersleben, G., Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. In: *Behavioral and Brain Sciences* 24, H. 5, S. 849–878.
- Kimänen, A., Alisaari, J., & Kallioniemi, A. (2019). In-service and pre-service teachers' orientations to linguistic, cultural and worldview diversity. In: *Journal of Teacher Education and Educators*, 8 (1), 35–54.
- Köker, A., Rosenbrock-Agyei, S., Ohm, U., Carlson, S. A., Ehmke, T., Hammer, S., Koch-Priewe, B., & Schulze, N. (2015). DaZKom – Ein Modell von Lehrerkompetenz im Bereich Deutsch als Zweitsprache. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Eds.): *Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie früh-pädagogischer Fachkräfte* (pp. 177–206). Bad Heilbrunn: Klinkhardt.
- Lemmrich, S., Bahls, A., & Ehmke, T. (under review). Effekte von mündlichen versus schriftlichen Antwortformaten bei der performanznahen Messung von Deutsch-als-Zweitsprache (DaZ)-Kompetenz – eine experimentelle Studie mit angehenden Lehrkräften.
- Lemmrich, S., Hecker, S.-L., Klein, S., Ehmke, T. Köker, A., Koch-Priewe, B., & Ohm, U. (2019). Performanznahe und videobasierte Messung von DaZ-Kompetenz bei Lehrkräften. Skalierung und dimensionale Struktur des Testinstruments. In T. Ehmke, P. Kuhl & M. Pietsch (Eds.), *Lehrer. Bildung. Gestalten. Beiträge zur empirischen Forschung in der Lehrerbildung* (pp. 188–202). Weinheim und Basel: Beltz Juventa.
- Lindmeier, A. (2013). Video-vignettenbasierte standardisierte Erhebung von Lehrerkognitionen. In U. Riegel & K. Macha (Eds.), *Videobasierte Kompetenzforschung in den Fachdidaktiken. Fachdidaktische Forschungen, Vol. 4*, pp. 45–62. Münster: Waxmann.
- Moosbrugger, H., & Kelava, A. (2007). Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer Medizin Verlag.
- Nimz, K., Hecker, S. – L., & Köker, A. (2018). Videobasierte Messung von DaZ-Kompetenz bei Lehrkräften. In C. Caruso, J. Hofmann, A. Rohde & K. Schick (Eds.), *Sprache im Unterricht. Ansätze, Konzept, Methoden* (pp. 439–452). Trier: WVT.
- Niesen, H. (2018). Förderung mehrsprachigkeitssensibler professioneller Handlungskompetenz angehender Englischlehrkräfte. In: *Zeitschrift für Interkulturellen Fremdsprachenunterricht: Didaktik und Methodik im Bereich Deutsch als Fremdsprache* 23, H. 1, S. 121–134.
- OECD (2005): PISA 2003. Technical report. Paris: OECD.

- Ohm, U. (2018). Das Modell von DaZ-Kompetenz bei angehenden Lehrkräften. In T. Ehmke, S. Hammer, A. Köker, U. Ohm & B. Koch-Priewe (Eds.), *Professionelle Kompetenzen angehender Lehrkräfte im Bereich Deutsch als Zweitsprache* (pp. 73–91). Münster: Waxmann.
- Rost, J. (2004). *Testtheorie Testkonstruktion* (second edition). Bern: Huber.
- Schmölzer-Eibinger, S. (2013). Sprache als Medium des Lernens im Fach. In M. Becker-Mrotzek, K. Schramm, E. Thürmann & H. J. Vollmer (Eds.), *Sprache im Fach. Sprachlichkeit und fachliches Lernen* (pp. 25–40). Münster: Waxmann.
- Seidel, T., Blomberg, G., & Stürmer, K. (2010). „Observer“ – Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. Projekt OBSERVE. In M. Bayrhuber, T. Leuders, R. Bruder & M. Wirtz (Eds.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (pp. 296–306). Weinheim und Basel: Beltz.
- Thürmann, E., & Vollmer, H. J. (2013). Schulsprache und Sprachsensibler Fachunterricht: Eine Checkliste mit Erläuterungen. In C. Röhner & B. Hövelbrinks (Eds.), *Fachbezogene Sprachförderung in Deutsch als Zweitsprache. Theoretische Konzepte und empirische Befunde zum Erwerb bildungssprachlicher Kompetenzen* (pp. 121–233). Weinheim: Beltz Juventa.
- van Es, E. A., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education*, 10 (4), pp. 571–596.
- Schleppegrell, M. J. (2004). *The language of school: A functional linguistics perspective*. Mahwah: Lawrence Erlbaum Associates.
- Weinert, F. E. (2001). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and Selecting Key Competencies* (pp. 45–66). Göttingen: Hogrefe.
- Wilson, M. (2005). *Constructing Measures. An Item Response Modeling Approach*. New York and London: Psychology Press, Taylor & Francis Group.



2.7

Effects of Early Childhood Teachers' Mathematics Anxiety on the Development of Childrens' Mathematical Competencies

Jenßen, L., Hosoya, G., Jegodtka, A., Eilerts, K., Eid, M., and Blömeke, S.

Abstract

Children develop their mathematical competencies already during their early years. Therefore, effective learning environments provided by early childhood teachers are required. Early childhood teachers' professional competence in mathematics is assumed to consist of different facets such as mathematical content knowledge and affective-motivational dispositions. Mathematics anxiety appears to be a common phenomenon amongst early childhood teachers and it is assumed that children educated by high math anxious teachers develop less mathematical competencies. To test this assumption, $n = 48$ early childhood teachers were tested regarding their mathematics anxiety and mathematical content knowledge and $n = 362$ corresponding children were tested twice within eight months regarding their mathematical competencies. Results indicate that children gain mathematical competencies over the eight-month period and that early childhood teachers' knowledge and anxiety in mathematics are negatively related. However, no effects of teachers' knowledge or anxiety on children's mathematical development were found. The discussion considers methodological implications and emphasizes limitations with respect to differences between the preschool context and primary or secondary school contexts.

Keywords

Early childhood teacher, mathematical development, early childhood education, mathematics anxiety, mathematical content knowledge

1 Introduction

The importance of early education in mathematics for children's competence development cannot be neglected and is of growing interest in research (Linder and Simpson 2017). Developing children's mathematical competence requires structured and systematic learning environments (Benz et al. 2016). Early Childhood Education and Care (ECEC) institutions intend to provide such mathematical learning environments; however, to succeed in this respect pedagogical professionals need mathematics-related knowledge regarding teaching mathematics, for example knowledge about designing effective learning environments, as well as regarding mathematical content domains (Tsamir et al. 2014; Bents et al. 2016; Dunekacke et al. 2016).

Studies revealed positive effects of ECEC institutions on the cognitive development of children in mathematics (Ulferts and Anders 2016). However, the role of ECEC teachers' emotions and to what extent they may promote or hinder children's competence development has not yet been examined. Emotions have an impact on learning mathematics in various ways (Schukajlow et al. 2017): Emotions directly and indirectly affect (mathematical) learning (Pekrun 2006), they mediate the transmission of mathematics knowledge to everyday life situations (Jansen et al. 2016), and they influence teachers' instructional strategies (Frenzel et al. 2017). Affective-motivational dispositions are conceptualized as inherent parts of early childhood teachers' professional competence (e.g., Tsamir et al. 2014). However, early childhood teachers' emotions related to mathematics and their effects on children's learning of mathematics have yet to be examined. This is the purpose of the present longitudinal study carried out in the Pro-KomMa project.¹

1 Pro-KomMa (Professionalization of Early Childhood Teacher Education) is a collaborative research project of Humboldt-Universität zu Berlin, Freie Universität Berlin and Alice Salomon University of Applied Sciences within the German research program Modeling and Measuring Competencies in Higher Education – Validation and Methodological Innovations (KoKoHs), funded by the Federal Ministry of Education and Research (BMBF).

2 Children's Mathematical Development in ECEC

Children start to acquire mathematical competence long before entering primary school (Benz et al. 2016). This mathematical development is assumed to happen constantly and children differ substantially at the end of preschool in their mathematical competence (Gervasoni and Perry 2015). Early mathematical competencies in the content domain “numbers and quantity” strongly predict later achievement in mathematics (Krajewski and Schneider 2009; Nguyen et al. 2016). Although a variety of mathematical content domains can be regarded as relevant for ECEC (Clements and Sarama 2011), theories and research have therefore strongly focused on the domain “numbers and quantity” (Dunckacke et al. 2018).

These competencies are characterized through “logical operations”, such as one-to-one assignments, invariance or part-whole relationships (Piaget and Szeminska 1975), a developing concept of numbers, for instance, reading and writing numbers (Clements and Sarama 2011) or, for example, through specific counting principles such as the cardinal principle (Gelman and Gallistel 1978). In particular, the link between numbers and quantities through a developing mathematical set is assumed to be a crucial factor (e.g., Resnick 1989). Consequently, current models, like the model by Krajewski regarding the development of the number concept, highlight the cognitive connection between quantities and numbers for children's mathematical development (Krajewski and Schneider 2009). With respect to this model, children show firstly basic mathematical competencies like seeing differences in size or knowing numerals. In the next stage, children link numerals with quantities (representation of sets) and relate quantities (e.g., part-whole relations) and, finally, they are able to compose and decompose numbers or form differences of specific numbers. The Pro-KomMa assessment of children's development in mathematics follows this model (Section 3.5).

Studies about learning mathematics through ECEC emphasize the importance of interactions between early childhood teacher and child (Doctoroff et al. 2016; Trawick-Smith et al. 2015). In particular, early childhood teachers' professional competence is expected to be highly important for children's acquisition of mathematics (Tirosh et al. 2011).

3 Early Childhood Teachers' Professional Competence

Professional competence can be conceptualized as a conglomerate of different dimensions: cognitive dispositions (knowledge), affective-motivational dispositions (e.g. beliefs and emotions), situation-specific skills (e.g. perception, interpretation,

decision-making), and observable performance (Blömeke et al. 2015a). Specific models concerning early childhood teachers also claim these different facets. Tsamir et al. (2014) highlight affective-motivational orientations (math-related self-efficacy) besides cognitive dispositions as the core of early childhood teachers' professional competence. Gasteiger and Benz (2018) propose a complex model, which can be considered an application and advancement of the competence conceptualization of Blömeke et al. (2015a). Besides different kinds of knowledge facets, the authors describe a cluster of attitudes, beliefs, and motivational orientations; a concrete description and operationalization, however, lacks. Other frameworks highlight performance-orientated competencies of early childhood teachers besides knowledge (e.g. Ginsburg et al. 2008; Clements and Sarama 2011).

The Pro-KomMa framework that underlies the present study can be situated alongside the models by Blömeke et al. (2015a) and Gasteiger and Benz (2018). One of its major assumptions is a significant relation between early childhood teachers' affective-motivational dispositions (e.g. math-related emotions, such as anxiety) and their mathematics-related knowledge. Another major assumption is that these in turn affect the development of children's mathematical competence (Jenßen et al. 2016).

3.1 Mathematics Anxiety

Mathematics anxiety can be described as a negative emotional response to math-related requirements in a variety of situations, such as in daily life or in academic situations (Cooke et al. 2011), which in turn leads to low achievement in mathematics (Ma 1999). Thus, it can be understood as an achievement emotion (Pekrun 2006). Mathematics anxiety is assumed to consist of different facets: The affective facets consist of feelings of fear of mathematics (primary emotional response) and feelings of helplessness, anger, and shame (secondary emotional response) (Buxton 1982; Cherkas 1992). The cognitive facet can be described as beliefs about one's own failure in mathematics and about mental blocks when working on math-related tasks (Hunt et al. 2014).

People who suffer from mathematics anxiety often report of physiological symptoms such as inner tension, transpiration, and higher muscle tonicity (Hembree 1990). In sum, mathematics anxiety seems to be a negative state that one wants to avoid. Consequently, the typical behavioral tendency coinciding with mathematics anxiety is the avoidance of math-related situations (Chang and Beilock 2016). People who report higher levels of mathematics anxiety also tend to avoid math-related tasks (Chipman et al. 1992). Although there is a lack of re-

search directly examining such consequences of math anxiety in ECEC teacher, studies reveal that early childhood teachers show lower self-confidence in creating mathematics learning environments and avoid math-related situations in the daily-preschool context (Bates et al. 2013).

Biological (e.g. genetics), psychological (e.g. introverted personality), social (e.g. negative parenting style), and pedagogical factors (e.g. lack of adequate mathematics education or right-or-wrong conceptions of mathematics at school) determine the development of mathematics anxiety (Ramirez et al. 2018). In accordance with Pekrun's Control-Value theory (Pekrun 2006), it can be assumed, that mathematics anxiety develops in light of two main appraisals: The value appraisal describes the importance of a domain, in this case mathematics, while the control appraisal describes the resources in a domain needed to meet a challenge (mainly cognitive resources, e.g. knowledge to solve mathematical tasks).

According to Benz (2012) and Thiel (2010), the majority of early childhood teachers in Germany reports positive beliefs regarding the importance of mathematics for everyday life and early education. However, positive beliefs about its importance do not necessarily result in positive emotions regarding mathematics. Positive beliefs about the importance of mathematics (value appraisal in terms of Control-Value theory) while believing in one's own low mathematical competence (control appraisal in terms of Control-Value theory) can in contrast result in mathematics anxiety.

It is important to consider that the majority of early childhood teachers are female, and mathematics anxiety is generally more prevalent in females than in males (Stoet et al. 2016). Mathematics anxiety is a common phenomenon in early childhood teachers (Bates et al. 2013; Gresham 2008; Gresham and Burleigh 2018). Besides primary school teachers, early childhood teachers describe themselves more often as anxious about mathematics than other pedagogical professionals do (Ginet et al. 2018). Younger pre-service early childhood teachers especially report more often to be anxious about mathematics (Thiel and Jenßen 2018). However, mathematics anxiety seems to not differ between pre-service and in-service early childhood teachers during their career entry (Gresham 2018). In contrary, some studies reveal that in-service early childhood teachers report even higher levels of mathematics anxiety (Aslan 2013). With respect to potential consequences, pre-service early childhood teachers with higher levels of mathematics anxiety show lower content knowledge in all mathematical domains (Jenßen et al. 2015a; Thiel and Jenßen 2018).

3.2 Early Childhood Teachers' Mathematics Content Knowledge

According to Shulman (1986), the cognitive disposition of teachers' professional competence can be differentiated into mathematics content knowledge (MCK), mathematics pedagogical content knowledge (MPCK), and general pedagogical knowledge (GPK). Usually, research on ECEC in the field of mathematics focuses on MPCK. However, an increasing number of studies revealed that MCK is important for perceiving situations from a mathematical point of view, for planning math-related activities (Dunekacke et al. 2015), and for performing adequately in math-related situations (Ginsburg and Ertle 2008; Klibanoff et al. 2006; see also Kaiser and König; and Kuhn et al. in this volume).

In theories and research concerning early education in mathematics, there is a long-standing debate about the nature of MCK needed for fostering children's mathematical competence (Ginsburg and Ertle 2008; Gasteiger and Benz 2018). Descriptions of MCK reach from rather school-orientated definitions (e.g. Tsamir et al. 2014) to conceptualizations of early childhood teachers' implicit MCK (e.g. Gasteiger and Benz 2018). However, all frameworks emphasize the importance of MCK for children's mathematical development.

Bäckman and Attorps (2012) postulate the *awareness of content* in mathematics. Early childhood teachers like other pedagogical professionals in mathematics have to know mathematics in depth and breadth. Depth of MCK describes the cumulative nature of mathematics, while breadth of MCK describes knowledge about the different mathematical domains. From the perspective of breadth, MCK can be described along the four mathematical domains "numbers, sets, and operations", "shape, space, and change", "quantity, measurement, and relations", and "data, combinatorics, and chance". This conceptualization that also underlies the Pro-KomMa project seems to be internationally valid, when comparing curricula from Germany and the U.S., for example (Jenßen et al. 2013).

However, regarding depth MCK seems to have a low value during early childhood teachers' training in Germany (Blömeke et al. 2017). As mathematics anxiety and MCK interact, reciprocal effects can be assumed (Carey et al. 2016): lower MCK may result in higher anxiety towards mathematics (and vice versa). Due to the hypothesized interaction of MCK and math anxiety, it is necessary to control for their MCK to analyze effects of early childhood teachers' mathematics anxiety.

3.3 Effects of Early Childhood Teachers' Math Anxiety and MCK on Children's Development

Research on mathematics anxiety highlights the transmission of adults' (e.g. parents or teachers) math anxiety to children (Herts et al. 2019). Consequences of teachers' mathematics anxiety for students are broadly discussed. High levels of mathematics anxiety are assumed to lead to high levels of mathematics anxiety and lower levels of mathematics knowledge in their students (Bekdemir 2010; Chang and Beilock 2016; O'Leary et al. 2017). In particular, transmissive models concerning emotions of teachers via their corresponding instructional behavior to their students, such as less process-orientated teaching strategies, might be an explanatory factor as well as children's perception of teachers' negative beliefs regarding the nature of mathematics (Hadley and Dorward 2011; Ramirez et al. 2018; Herts et al. 2019; 2003Frenzel et al. 2017). The transmissive power of math-specific beliefs or gender stereotypes in adult-child-interactions may be also explanatory factors (Herts et al. 2019).

So far, the assumed link between teachers' high level of mathematics anxiety and students' low level of mathematics achievement could only be validated empirically for primary and secondary school teachers (Beilock et al. 2010; Hadley and Dorward 2011; Ramirez et al. 2018) their math anxiety carries negative consequences for the math achievement of their female students. Early elementary school teachers in the United States are almost exclusively female (>90 %). Within the context of ECEC, many factors regarding early childhood teachers due to higher math anxiety could be theoretically assumed as explanations for children's potentially lower achievement in mathematics: Negation of the importance of mathematics in ECEC, underestimation of mathematics activities in preschool, resulting in lower mathematics content knowledge, avoidance of math-related situations and/or transmission of negative thoughts and feelings in interactions. The opportunity for avoidance of math-related situations might be particularly prevalent within the informal learning context of ECEC. With respect to early childhood teachers, with the exception of one study that is characterized by methodological problems (Aslan et al. 2013), no empirical validation of a transmission assumption regarding math anxiety exists. Our longitudinal study is the first of its kind.

While effects of teachers' MCK on children's achievement can be found for primary school (e.g. Hill et al. 2005), empirical studies for early childhood teachers are lacking, with the exception of a few qualitative studies (Tirosch et al. 2011). Although limited generalization these studies indicate that it might be an advantage for children's mathematical development if their ECEC teachers dispose of rich


MCK. Although MCK mainly serves as a control variable in our study, it will be possible to examine this relationship as a side effect.

3.4 Research Questions and Hypothesis

The present longitudinal study investigates whether early childhood teachers' mathematics anxiety shows effects on children's mathematical development. Teachers' MCK is included in the model to control for cognitive dispositions of professional competence. In line with the transmission theory, we hypothesize that ECEC teachers' mathematics anxiety has a negative effect on children's mathematical development (H1). Derived from theoretical assumptions and empirical studies, we hypothesize in contrast that their MCK has a positive effect (H2). As indicated by the state of research, we assume a negative relation of medium size between mathematics anxiety and MCK (H3). As research on direct effects of early childhood teachers' emotion on children's mathematical development is scarce, we refrain from hypothesizing the size of potential effects.

3.5 Instruments

To assess early childhood teachers' MCK, the KomMa-MCK-Test (Blömeke et al. 2015b) was used. The test consists of 24 items covering four mathematical domains of early education (numbers, sets, and operations; shape, space, and change; quantity, measurement, and relations; data, combinatorics, and chance). Each domain consists of six items in multiple-choice and open-response format. The test has been validated in multiple studies regarding content (Jenßen et al. 2015b), structure (Blömeke et al. 2015b), and relation to other variables, such as learning opportunities during early educators' training (Blömeke et al. 2017), beliefs (Dunekacke et al. 2016), situation-specific perception (Dunekacke et al. 2015), and general cognitive abilities (Blömeke and Jenßen 2017). Figure 1 shows an example of a multiple-choice item concerning the domain "data, combinatorics, and chance".



Chris has a blue, a green, a red and a yellow cube.
Chris wants to pile up a tower with the four cubes

Which arithmetic expression provides the number of possibilities of the different towers?

Please indicate your answer with a cross

- $4 + 4 + 4 + 4$
- $4 \cdot 4 \cdot 4 \cdot 4$
- $4 \cdot 3 \cdot 2$
- $4 + 3 + 2$

Figure 1 Example item from the KomMa-MCK-Test (translated)

Mathematics anxiety was assessed using the Mathematics Anxiety Scale – Revised (MAS-R) (Bai et al. 2009). The questionnaire contains 14 items of which six are positive statements (e.g. “I find math interesting”) and eight are negative statements (e.g. “Mathematics makes me feel nervous.”). Answers had to be rated on a five-point Likert scale ranging from “totally agree” to “totally disagree.” Positive statements have to be inverted so high scores indicate higher anxiety towards mathematics. Bai et al. state that negative statements of the assessment reflect the affective facet of mathematics anxiety and the (later inverted) positive statement reflect the cognitive facet. In the past, MAS-R was used to assess pre-service early childhood teachers’ mathematics anxiety in different research contexts including Germany (Jenßen et al. 2015a; Thiel and Jenßen 2018) and can be considered valid concerning relations to other variables and factorial structure (Bai et al. 2009).

A standardized assessment well-established in Germany was used to test children’s mathematical achievement (Test mathematischer Basiskompetenzen im Kindergarten, MBK 0; Krajewski 2018). The test is based on Krajewski’s model of the development of children’s quantity-number competencies (Krajewski and Schneider 2009) and considers three major areas: number meaning, sets of numbers,

and number relations. Tasks contain typical procedures of children's mathematical activities such as counting forward, naming numerals, ordering numbers, and comparing quantities. The test can be predominantly applied for below average and average areas on the achievement continuum.

3.6 Participants and Procedure

$N = 362$ children were tested twice within approximately 8 months in their regular preschool context. The children's mean age at the first time-point of assessment on the MBK 0 was $M = 1633$ days, respectively 4.5 years, ($SD = 311$ days) and 56% of the participating children were girls. The MBK 0 was administered by two trained university students. About 8 children were grouped with one early childhood teacher ($M = 7.5$), whereby only one early childhood teacher was related to one preschool, respectively to one group of children.

About 1.5 months before the second test administration, $n = 48$ early childhood teachers' MCK and mathematics anxiety were assessed via an online-based assessment. The teachers' mean age was $M = 33$ years ($SD = 9.3$). One item of the KomMa-MCK-Test concerning "shape, space, and change" had to be excluded for online assessment because of technical difficulty (the open-response item required the drawing of a cuboid, which was not technically possible). The majority of the early childhood teachers were female (91.7%) and all teachers were career entrants with five years' experience at a maximum. Teachers received monetary incentives for their participation.

3.7 Data Analysis

A two-level change model for two measurement occasions was specified for data analysis. In Figure 2, (intercepts within) represents the children's scores on the MBK 0 on the first measurement occasion and (slopes within) represents the children's change on the MBK 0 from time-point 1 to time-point 2. As children are nested within teachers, (intercepts between) represents average scores of the children within a group at the first measurement occasion and (slopes between) represents the group-specific change of the children's performance on the teacher level.

The group-specific change of the children's performance is regressed on math anxiety as well as on MCK which were assessed on the teacher level. Both variables were standardized on the teacher level in the analysis to allow for a comparison of the effects on the group-specific change and to provide a meaningful

interpretation of the intercepts on the teacher level. In addition, the covariance between MCK and math anxiety were modelled.

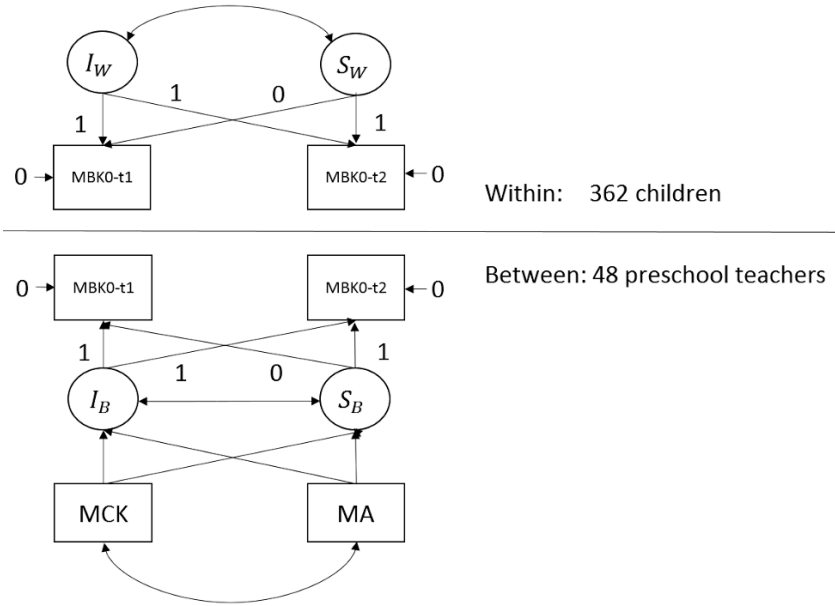


Figure 2 Two-level change model for two measurement occasions

$MBK0-t_1$: MBK 0 scores on the first measurement occasion. $MBK0-t_2$: MBK 0 scores on the second measurement occasion. MCK: Teachers’ mathematical content knowledge. MA: Teacher’s math anxiety.

The data were analyzed with the software MPlus (Muthén and Muthén 1998–2017), version 8.3. As the data contained a large amount of missing values, only 126 cases contained no missings (Table 1), full information maximum likelihood estimation (FIML) was used. In addition, a sensitivity analysis was conducted using multiple imputation using the unrestricted H1 model to control if the results were stable across different analysis conditions. 100 Imputation datasets and robust maximum likelihood estimation (MLR) were used.

Table 1 Frequencies (*n*) of missingness patterns in the data set (*N* = 362)

	Pattern 1 (<i>n</i> = 126)	Pattern 2 (<i>n</i> = 82)	Pattern 3 (<i>n</i> = 56)	Pattern 4 (<i>n</i> = 46)	Pattern 5 (<i>n</i> = 19)	Pattern 6 (<i>n</i> = 10)	Pattern 7 (<i>n</i> = 8)	Pattern 8 (<i>n</i> = 6)	Pattern 9 (<i>n</i> = 7)	Pattern 10 (<i>n</i> = 2)
MAS-R	X	X			X		X			
MCK	X	X			X		X	X	X	X
MBK 0 t ₁	X	X	X	X				X	X	
MBK 0 t ₂	X			X	X	x			X	X

Note. X: not missing, MAS-R: Math Anxiety Scale Revised; MCK: Mathematical Content Knowledge, MBK 0 t₁; Children's competence on first measurement occasion; MBK 0 t₂; Children's competence on second measurement occasion.

4 Results

For $n = 30$ early childhood teachers MCK scores were available. Scores of the KomMa-MCK-Test can range from 0 to 23 points. On average, teachers reached 12.9 points ($SD = 6.04$, $min = 1$, $max = 22$). The MAS-R scores can range from 14 to 70 points. Mathematics anxiety scores were available for $n = 28$ teachers ($M = 40.4$, $SD = 11.00$, $min = 16$, $max = 61$).

Table 2 shows the results of the analysis. With FIML-estimation, the estimated average score of the children on the MBK 0 at the first measurement occasion is $= 14.4$ ($p < .001$). On average, the children's groups gained 8 points on the MBK 0 from measurement occasion 1 to measurement occasion 2 ($p < .001$). The minimum score achievable on the MBK 0 is 0 points, the maximum score is 44 points.

When using FIML estimation, the estimated effect of math anxiety on the children's groups' gain-scores is negative and statistically significant ($= 1.4$; $p = .04$), as hypothesized (H1): with a self-reported increase in ECEC teachers' math anxiety of one standard deviation (11 points on the math anxiety scale), their children group's gains were reduced by expected 1.4 points compared to a group with an ECEC teacher of similar MCK.

In contrast to our hypothesis (H2), there was no significant effect of MCK on the children's groups' development ($= -1.128$; $p = .198$) when controlling for math anxiety. Together, math anxiety and MCK explain 28.6% ($p = .181$) of the variance in the children's groups' gain scores. However, the variance explained is statistically not significant.

In line with our hypothesis (H3), the covariance of math anxiety with MCK on the between level is negative ($= -0.51$; $p < .028$) and statistically significant, indicating that teachers with a high level of math anxiety are more likely to have lower MCK. The respective estimated correlation between math anxiety and MCK is $= -.50$ and thus quite strong.

The estimates using multiple imputation are relatively similar to those obtained by FIML estimation. However, the estimated effect of math anxiety on the children's performance is statistically not significant ($= -1.1$; $p = .265$) (H1). The estimated effects for the expected average gain of the children on the MBK 0 ($= 7.7$; $p < .001$) and the negative covariance between math anxiety and MCK ($= -0.49$; $p = .048$) with a respective correlation of $= -.47$ remain robust across analysis methods. One reason for the divergence between the FIML- and multiple imputation analyses could lie on a large amount of missing data. Multiple imputation considers the uncertainty induced by the missing data and, thus, the method may result in slightly different estimates as well as, in some cases, slightly larger standard errors as compared to FIML estimation.

Table 2 Estimated unstandardized model coefficients for full information maximum likelihood (FIML) and multiple imputation (MI)

	FIML-Esti- mate (SE)	FIML p-value	MI-Esti- mate (SE)	MI p-Value	Rate of Missings
Within Level					
SW with IW	-18.099 (4.803)	<.001	-19.168 (5.805)	<.01	.556
Variances					
IW	79.174 (8.991)	<.001	78.413 (8.364)	<.001	.077
SW	47.234 (5.430)	<.001	49.642 (6.265)	<.001	.575
Between Level					
IB on Math Anxiety	-0.118 (1.766)	.947	0.625 (2.068)	.762	.566
IB on MCK	1.743 (2.261)	.441	2.037 (2.065)	.324	.477
SB on Math Anxiety	-1.389 (0.687)	.043	-1.083 (0.972)	.265	.486
SB on MCK	-1.128 (0.876)	.198	-0.793 (1.069)	.458	.536
SB with IB	-4.209 (6.440)	.513	-5.050 (7.532)	.503	.405
Math Anxiety with MCK	-0.506 (0.230)	.028	-0.490 (0.248)	.048	.547
Means					
Math Anxiety	0.095 (0.186)	.612	0.070 (0.193)	.716	.436
MCK	-0.003 (0.196)	.987	-0.005 (0.184)	.977	.352
Intercepts					
IB	14.425 (1.372)	<.001	14.521 (1.390)	<0.001	.063
SB	7.951 (0.626)	<.001	7.715 (0.804)	<0.001	.419
Variances					
Math Anxiety	0.995 (0.261)	<.001	1.014 (0.320)	<0.01	.605
MCK	1.035 (0.186)	<.001	1.053 (0.251)	<0.001	.480

	FIML-Estimate (SE)	FIML p-value	MI-Estimate (SE)	MI p-Value	Rate of Missings
Residual Variances					
IB	79.998 (16.207)	<.001	74.720 (15.618)	<.001	.161
SB	4.130 (3.501)	.238	10.763 (6.246)	0.085	.532

Note. IW: Intercepts within; SW: Slopes within; IB: Intercepts between; SB: Slopes between. MCK: Mathematical Content Knowledge; Effects of interest are printed in bold font. MCK and math anxiety were standardized on the between level to allow for a comparison of the effects.

5 Discussion and Conclusion

Our study contributes to other findings, that children gain mathematical competencies within the ECEC institutions in Germany (Ulferts and Anders 2016). As hypothesized based on results in other studies, early childhood teachers' math anxiety and MCK are negatively associated. However, they seem to be stronger related compared to studies with pre-service early childhood teachers (e.g. Jenßen et al. 2015a; Thiel and Jenßen 2018). These results are robust across analysis methods.

In this study, no conclusive links between early childhood teachers' characteristics and the development of the children could be established though. Albeit the effect of math anxiety on the children's development is statistically significant using FIML estimation, multiple imputation paints a slightly different picture. The fact that the amount of variance in the group specific gain scores explained by the teacher characteristics is not significant using FIML estimation also stresses the conclusion that no decisive link between teacher characteristics and the children's development could be established. Thus, our study replicates findings by Aslan et al. (2013). However, also the study by Aslan et al. (2013) was characterized by methodological problems such as the lacking consideration of clustered data structure so that further studies are needed to conclusively test a possible link between teacher characteristics and children's development.

In contrast to our hypothesis, early childhood teachers' MCK had no effect on children's mathematical development in the present study. The state of research was already inconclusive in this respect. One explanation could be our conceptualization of early childhood teachers' MCK, which could be described as more orientated towards the transition from ECEC towards primary school and thus towards slightly older children (Blömeke et al. 2015b). Hedges and Cullen (2006) emphasize the importance of early childhood teachers' MCK for children's math-

emathical development. From a conceptual point of view, they state that early childhood teachers' MCK should be related to children's mathematical knowledge that is needed for mathematical development and for later school achievement. Thus, our conceptualization should reveal an effect in the present study as well but from an empirical point of view, it is still unknown which kind of MCK early childhood teachers need to foster children's mathematical competence. It could be one of the most important desiderata in ECEC research in mathematics, to clarify whether more "child-centered" MCK (respectively basic mathematical concepts such as declarative MCK) or more "school-orientated" MCK (respectively advanced mathematical concepts such as arithmetic MCK in the sense of school conceptualizations) is needed.

Another explanation could be that we treated the link between MCK and children's development as a "black box". Data is needed about other facets of ECEC teachers' professional competence and their actual behavior in the classroom as well as about other contextual characteristics of children's learning environment including their opportunities to learn mathematics at home. Especially gender-related stereotype beliefs seem to play an important role in the transmission of math anxiety (Herts et al. 2019), but empirical research is still lacking. In this case, it would be possible to specify more complex models that take a broader range of predictors and potentially mediating factors into account. This would require a large range of measures though while it is questionable whether early childhood teachers would take a test battery that lasts for several hours besides their job duties.

There are also some limitations of our study. The representativeness of the teacher sample regarding job experience, gender, and regional specifics during teachers' training in Germany is limited as the participating early childhood teachers were at the beginning of their careers (although some of the early childhood teachers seem to have completed a former different training), mainly female, and were working in Berlin and Brandenburg. In particular, gender seems to be a highly important feature in math (Beilock et al. 2010) their math anxiety carries negative consequences for the math achievement of their female students. Early elementary school teachers in the United States are almost exclusively female (>90 %, especially in the context of ECEC, because the majority of early childhood teachers in Germany are female. Opportunities to learn in mathematics during early childhood teachers' training may also be a limiting factor (Blömeke et al. 2017). In 2012, there has been a major turn in the system of vocational training for early childhood teachers in Germany. Since then, there has been a transnational framework curriculum, which has been implemented in all federal states of Germany. In this framework, natural science education including mathematical and technical

education has been defined as one main competence area in ECEC. Thus, early childhood teachers have the opportunity during their training to deal with STEM topics such as mathematics. However, mathematics is still not an exclusive competence area besides others in this framework, but only a subdomain. This aspect may be a specific one for ECEC teachers' training in Germany. It could be also important, that Germany shows a more social-educational approach in ECEC as compared to other countries such as the U.S. Moreover, in the current application only one early childhood teacher per group was tested. Usually, there are at least two early childhood teachers responsible for a group of children. From this structural point of view, preschool differs from primary school. Thus, our assumption of the impact of *one* early childhood teacher's math anxiety on children's mathematical development derived from theories based on primary or secondary school structures may be more or less applicable. Additionally, it has to be assumed that learning groups in preschool are more heterogeneous than in primary or secondary school. Acquisition of cognitive competencies or transmission of teachers' beliefs and/or emotions may be less powerful in these heterogeneous groups. Future research should consider these points.

To conclude, although it remains unclear whether math anxiety shows direct effects on children's learning in the ECEC context in Germany, math anxiety seems to be a relevant dispositional facet of early childhood teachers. In itself this is an unfavorable characteristic. Consequently, the reduction of early childhood teachers' mathematics anxiety seems to be a goal for training and practice. A promising intervention for early childhood teachers' mathematics anxiety may be the reduction of the avoidance of math-related situations. Firstly, early childhood teachers have to be aware of their own level of mathematics anxiety. According to Polya (1977), it is mandatory to experience mathematics as a domain full of positive aspects, so that the teacher gains positive feelings and thoughts concerning mathematics. Secondly, mathematics has to be valued as an important subject by early childhood teachers within the ECEC context. Thirdly, with respect to Control-Value Theory (Pekrun 2006), it is necessary for early childhood teachers to think positively about their own resources as a way of coping with challenging mathematical tasks in ECEC situations. Early childhood teachers need to develop their own MCK during training (Blömeke et al. 2017), so that they feel competent in mathematics (Thiel and Jenßen 2018).

References

- Aslan, D. (2013). A comparison of pre- and in-service preschool teachers' mathematical anxiety and beliefs about mathematics for young children. *Academic Research International*, 4(2), 225–230. Retrieved from [http://www.savap.org.pk/journals/ARInt/Vol.4\(2\)/2013\(4.2-22\).pdf](http://www.savap.org.pk/journals/ARInt/Vol.4(2)/2013(4.2-22).pdf)
- Aslan, D., Oğul, İ. G., & Taş, I. (2013). The Impacts of Preschool Teachers' Mathematics Anxiety and Beliefs on Children's Mathematics Achievement. *International Journal of Humanities and Social Science Invention*, 2(7), pp. 45–49.
- Bai, H., Wang, L., Pan, W., & Frey, M. (2009). Measuring mathematics anxiety: Psychometric analysis of a bidimensional affective scale. *Journal of Instructional Psychology*, 36(3), pp. 189–193. Retrieved from <http://psycnet.apa.org/psycinfo/2009-20155-001>
- Bäckman, K., & Attorps, I. (2012). Teaching mathematics in the pre-school context. *US-China Education Review*, pp. 1–16.
- Bates, A. B., Latham, N. I., & Kim, J. (2013). Do I Have to Teach Math? Early Childhood Pre-Service Teachers' Fears of Teaching Mathematics. *IUMPST: The Journal*, 5(August).
- Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences*, 107(5), pp. 1860–1863. <https://doi.org/10.1073/pnas.0910967107>
- Bekdemir, M. (2010). The pre-service teachers' mathematics anxiety related to depth of negative experiences in mathematics classroom while they were students. *Educational Studies in Mathematics*, 75(3), pp. 311–328. <https://doi.org/10.1007/s10649-010-9260-7>
- Benz, C. (2012). "Maths is not dangerous" – Attitudes of people working in German kindergarten about mathematics in kindergarten". *European Early Childhood Education Journal*, 20(2), pp. 249–261. doi:10.1080/1350293X.2012.681131
- Benz, C., Steinweg, A.S., Gasteiger, H., Schöner, P., Vollmuth, H., & Zöllner, J. (2018). *Mathematics education in the early years*. Cham, Switzerland: Springer International.
- Blömeke, S., & Jenßen, L. (2017). A question of validity: Clarifying the hierarchical nature of teacher cognition. In M. Rosén, K. Y. Hansen, & U. Wolff (Eds.), *Cognitive abilities and educational outcomes. A festschrift in honour of Jan-Eric Gustafsson* (pp. 89–110). Cham, Switzerland: Springer International Publishing.
- Blömeke, S., Jenßen, L., Grassmann, M., Dunekacke, S., & Wedekind, H. (2017). Process Mediates Structure: The Relation Between Preschool Teacher's Education and Preschool Teachers' Knowledge. *Journal of Educational Psychology*, 109(3), pp. 338–354. <https://doi.org/10.1037/edu0000147>
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015a). Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), pp. 3–13. <https://doi.org/10.1027/2151-2604/a000194>
- Blömeke, S., Jenßen, L., Dunekacke, S., Suhl, U., Grassmann, M., & Wedekind, H. (2015b). Leistungstests zur Messung der professionellen Kompetenz frühpädagogischer Fachkräfte. *Zeitschrift Für Pädagogische Psychologie*, 29(3–4), pp. 177–191. <https://doi.org/10.1024/1010-0652/a000159>
- Carey, E., Hill, F., Devine, A., & Szűcs, D. (2016). The chicken or the egg? The direction of the relationship between mathematics anxiety and mathematics performance. *Frontiers in Psychology*, 6, pp. 1–6. <https://doi.org/10.3389/fpsyg.2015.01987>

- Chang, H., & Beilock, S. L. (2016). The math anxiety-math performance link and its relation to individual and environmental factors: A review of current behavioral and psychophysiological research. *Current Opinion in Behavioral Sciences*, 10, pp. 33–38. <https://doi.org/10.1016/j.cobeha.2016.04.011>
- Chipman, S. F., Krantz, D. H., & Silver, R. (1992). Mathematics Anxiety and Science Careers Among Able College-Women. *Psychological Science*, 3(5), pp. 292–295. <https://doi.org/10.1111/j.1467-9280.1992.tb00675.x>
- Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, 333, pp. 968–970. doi: 10.1126/science.1204537
- Cooke, A., Cavanagh, R., Hurst, C., & Sparrow, L. (2011). *Situational effects of mathematics anxiety in pre-service teacher education*. Paper presented at the 2011 AARE international Research Conference, Hobart, Australia, 27 November–1 December, 2011. Retrieved from <http://www.aare.edu.au/data/publications/2011/aarefinal00501.pdf>
- Doctoroff, G. L., Fisher, P. H., Burrows, B. M., & Edman, M. T. (2016). Preschool children's interest, social-emotional skills, and emergent mathematics skills. *Psychology in the Schools*, 53(4), pp. 390–403. <https://doi.org/10.1002/pits>
- Dunekacke, S., Grüßing, M., & Heinze, A. (2018). Is considering numerical competence sufficient? The structure of 6-year-old preschool children's mathematical competence. In C. Benz, A.S. Steinweg, H. Gasteiger, P. Schöner, H. Vollmuth, & J. Zöllner (Eds.), *Mathematics education in the early years* (pp. 145–157). Cham, Switzerland: Springer International.
- Dunekacke, S., Jenßen, L., Eilerts, K., & Blömeke, S. (2016). Epistemological beliefs of prospective preschool teachers and their relation to knowledge, perception, and planning abilities in the field of mathematics: a process model. *ZDM – Mathematics Education*, 48(1–2), pp. 125–137. <https://doi.org/10.1007/s11858-015-0711-6>
- Dunekacke, S., Jenßen, L., & Blömeke, S. (2015). Effects of Mathematics Content Knowledge on Pre-school Teachers' Performance: a Video-Based Assessment of Perception and Planning Abilities in Informal Learning Situations. *International Journal of Science and Mathematics Education*, 13(2), pp. 267–286. <https://doi.org/10.1007/s10763-014-9596-z>
- Frenzel, A. C., Becker-kurz, B., Pekrun, R., Goetz, T., Lüdtke, O., Frenzel, A. C., ... Lüdtke, O. (2017). Reciprocal effects model of teacher and student emotion transmission in the classroom revisited: A reciprocal effects model of teacher and student enjoyment. *Journal of Educational Psychology*, pp. 1–12.
- Gasteiger, H., & Benz, C. (2018). Mathematics education competence of professionals in early childhood education: A theory-based competence model. In C. Benz, A.S. Steinweg, H. Gasteiger, P. Schöner, H. Vollmuth, & J. Zöllner (Eds.), *Mathematics education in the early years* (pp. 69–92). Cham, Switzerland: Springer International.
- Gelman, R., & Gallistel, C.R. (1978). *The child's understanding of number*. Cambridge, MA: Harvard University Press.
- Gervasoni, A., & Perry, B. (2015). Children's mathematical knowledge prior to starting school and implications for transition. In B. Perry, A. MacDonald, & A. Gervasoni (Eds.), *Mathematics and transition to school* (pp. 47–64). Wiesbaden, Germany: Springer.
- Ginet, L., Itzkowich, R., & Maloney, E. (2018). Math anxiety and math performance: How do they relate? In J.S. McCray, J.-Q. Chen, & J.E. Sorkin (Eds.), *Growing mathematical minds. Conversations between developmental psychologists and early childhood teachers* (pp. 173–200). New York: Routledge.

- Ginsburg, H. P., & Ertle, B. (2008). Knowing the mathematics in early childhood mathematics. In O. N. Saracho & B. Spodek (Eds.), *Contemporary perspectives on mathematics in early childhood education* (pp. 45–66). Charlotte, NC: Information AGE.
- Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report – Giving Child and Youth Development Knowledge Away*, 22(1), pp. 1–24.
- Gresham, G. (2018). Preservice to Inservice: Does Mathematics Anxiety Change With Teaching Experience? *Journal of Teacher Education*, 69(1), pp. 90–107. <https://doi.org/10.1177/0022487117702580>
- Gresham, G. (2008). Mathematics anxiety and mathematics teacher efficacy in elementary pre-service teachers. *Teaching Education*, 19(3), pp. 171–184. <https://doi.org/10.1080/10476210802250133>
- Gresham, G., & Burleigh, C. (2018). Exploring early childhood preservice teachers' mathematics anxiety and mathematics efficacy beliefs. *Teaching Education*, pp. 1–25. <https://doi.org/10.1080/10476210.2018.1466875>
- Hadley, K. M., & Dorward, J. (2011). Investigating the Relationship between Elementary Teacher Mathematics Anxiety, Mathematics Instructional Practices, and Student Mathematics Achievement. *Journal of Curriculum and Instruction*, 5(2), pp. 27–44. <https://doi.org/10.3776/joci.2011.v5n2p27-44>
- Hedges, H., & Cullen, J. (2005). Meaningful teaching and learning: Children's and teachers' content knowledge. *ACE Paper: Approaches to Domain Knowledge in Early Childhood Pedagogy*, 16, pp. 11–24. Retrieved from <https://researchspace.auckland.ac.nz/handle/2292/25146>
- Hembree, R. (1990). The Nature, Effects, and Relief of Mathematics Anxiety. *Journal for Research in Mathematics Education*, 21(1), pp. 33–46.
- Herts, J.B., Beilock, S.L., & Levine, S.C. (2019). The role of parents' and teachers' math anxiety in children's math learning and attitudes. In I.C. Mammarella, S. Caviola, & A. Dowker (Eds.), *Mathematics anxiety: What is known and what is still to be understood* (pp.190–210). London & New York: Routledge.
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal*, 42(2), pp. 371–406. <https://doi.org/10.3102/00028312042002371>
- Hunt, T. E., Clark-carter, D., Sheffield, D., & Hunt, T. E. (2014). Math Anxiety, Intrusive Thoughts and Performance. *Journal of Education, Psychology and Social Sciences*, 2(2), pp. 69–75.
- Jansen, B. R. J., Schmitz, E. A., & van der Maas, H. L. J. (2016). Affective and motivational factors mediate the relation between math skills and use of math in everyday life. *Frontiers in Psychology*, 7(APR), pp. 1–11. <https://doi.org/10.3389/fpsyg.2016.00513>
- Jenßen, L., Jegodtka, A., Eilerts, K., Eid, M., Koinzer, T., Schmude, C., Rasche, J., Szczesny, M., & Blömeke, S. (2016). Pro-KomMa – Professionalization of Early Childhood Teacher Education: Convergent, Discriminant, and Prognostic Validation of the KomMa Models and Tests. In H.A. Pant, O. Zlatkin-Troitschanskaia, C. Lautenbach, M. Toepper, & D. Molerov (Eds.), *Modeling and Measuring Competencies in Higher Education – Validation and Methodological Innovations (KoKoHs) – Overview of the Research Projects* (pp. 39–43). Berlin & Mainz: Humboldt University & Johannes Gutenberg University.

- Jenßen, L., Dunekacke, S., Eid, M., & Blömeke, S. (2015a). The relationship of mathematical competence and mathematics anxiety: An application of latent state-trait theory. *Zeitschrift Fur Psychologie/Journal of Psychology*, 223(1), pp. 31–38. <https://doi.org/10.1027/2151-2604/a000197>
- Jenßen, L., Dunekacke, S., & Blömeke, S. (2015b). Qualitätssicherung in der Kompetenzforschung. *Kompetenzen von Studierenden*. 61. Beiheft Der Zeitschrift Für Pädagogik, 2015, pp. 11–31.
- Jenßen, L., Dunekacke, S., Baack, W., Tengler, M., Wedekind, H., Grassmann, M., & Blömeke, S. (2013). Validating an assessment of pre-school teachers' mathematical knowledge. *Paper presented at the meeting of the International Group for the Psychology of Mathematics Education*, Kiel, Germany.
- Klibanoff, R. S., Levine, S. C., Huttenlocher, J., Vasilyeva, M., & Hedges, L. V. (2006). Preschool children's mathematical knowledge: The effect of teacher "math talk." *Developmental Psychology*, 42(1), pp. 59–69. doi: 10.1037/0012-1649.42.1.59
- Krajewski, K. (2018). *MBK 0. Test mathematischer Basiskompetenzen im Kindergartenalter*. Göttingen: Hogrefe.
- Krajewski, K., & Schneider, W. (2009). Early development of quantity to number-word linkage as a precursor of mathematical school achievement and mathematical difficulties: Findings from a four-year longitudinal study. *Learning and Instruction*, 19(6), pp. 513–526.
- Linder, S. M., & Simpson, A. (2017). Towards an understanding of early childhood mathematics education: A systematic review of the literature focusing on practicing and prospective teachers. *Contemporary Issues in Early Childhood*, pp. 1–23. <https://doi.org/10.1177/1463949117719553>
- Ma, X. (1999). A Meta-Analysis of the Relationship between Anxiety toward Mathematics and Achievement in Mathematics. *Journal for Research in Mathematics Education*, 30(5), pp. 520–540.
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén.
- Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., & Spitler, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research Quarterly*, 26, pp. 550–560.
- O'Leary, K., Fitzpatrick, C. L., & Hallett, D. (2017). Math anxiety is related to some, but not all, experiences with math. *Frontiers in Psychology*, 8(DEC), pp. 1–14. <https://doi.org/10.3389/fpsyg.2017.02067>
- Peckun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), pp. 315–341.
- Piaget, J., & Szeminska, A. (1975). *Die Entwicklung des Zahlbegriffs beim Kinde*. Stuttgart: Klett-Cotta.
- Polya, G. 1977. *Mathematical methods in science*. Washington, DC: Mathematical Association of America.
- Ramirez, G., Hooper, S. Y., Kersting, N. B., Ferguson, R., & Yeager, D. (2018). Teacher Math Anxiety Relates to Adolescent Students' Math Achievement. *AERA Open*, 4(1), pp. 1–13. <https://doi.org/10.1177/2332858418756052>

- Ramirez, G., Shaw, S. T., & Maloney, E. A. (2018). Math anxiety: Past research, promising interventions, and a new interpretation framework. *Educational Psychologist, 1520*, pp. 1–20. <https://doi.org/10.1080/00461520.2018.1447384>
- Schukajlow, S., Rakoczy, K., & Pekrun, R. (2017). Emotions and motivation in mathematics education: theoretical considerations and empirical contributions. *ZDM – Mathematics Education, 49*(3), pp. 307–322. <https://doi.org/10.1007/s11858-017-0864-6>
- Stoet, G., Bailey, D. H., Moore, A. M., & Geary, D. C. (2016). Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. *PLoS ONE, 11*(4), pp. 1–24. <https://doi.org/10.1371/journal.pone.0153857>
- Thiel, O. (2010). Teachers' attitudes towards mathematics in early childhood education. *European Early Childhood Education Research Journal, 18*(1), pp. 105–115. doi: 10.1080/13502930903520090
- Thiel, O., & Jenßen, L. (2018). Affective-motivational aspects of early childhood teacher students' knowledge about mathematics. *European Early Childhood Education Research Journal, 26*(4), pp. 512–534. <https://doi.org/10.1080/1350293X.2018.1488398>
- Tirosh, D., Tsamir, P., Levenson, E., & Tabach, M. (2011). From preschool teachers' professional development to children's knowledge: comparing sets. *Journal of Mathematics Teacher Education, 14*(2), pp. 113–131. <https://doi.org/10.1007/s10857-011-9172-1>
- Trawick-Smith, J., Swaminathan, S., & Liu, X. (2015). The relationship of teacher-child play interactions to mathematics learning in preschool. *Early Child Development and Care, 186*(5), pp. 716–733. <https://doi.org/10.1080/03004430.2015.1054818>
- Tsamir, P., Tirosh, D., Levenson, E., Tabach, M., & Barkai, R. (2014a). Employing the CAMTE framework: Focusing on preschool teachers' knowledge and self-efficacy related to students' conceptions. In U. Kortenkamp, B. Brandt, C. Benz, G. Krummheuer, S. Ladel & R. Vogel (Eds.), *Early Mathematics Learning* (pp. 291–306). New York: Springer Science+Business.
- Ulferts, H., & Anders, Y. (2016). *Effects of ECEC on academic outcomes in literacy and mathematics: Meta-analysis of European longitudinal studies.*

Generic Competencies and Their Impact on Learning in Higher Education

3



3.1

Modelling, Assessing, and Promoting Competences for Self-Regulated Learning in Higher Education

Eckerlein, N., Dresel, M., Steuer, G., Foerst, N., Ziegler, A., Schmitz, B., Spiel, C., and Schober, B.

Abstract

Summarized are the results of a research project that focused on assessing and fostering competences for self-regulated learning in higher education. Innovative instruments (situational judgement test, competence-performance-assessment, learning journal) to assess different aspects of self-regulated learning (e.g., use of cognitive and metacognitive strategies, strategies of motivational regulation) were developed and validated. Moreover, training approaches to foster different aspects of self-regulated learning strategy use were conceptualized, also to test the change sensitivity of the assessments. In detail, the importance of competences for self-regulated learning for central components of successful learning processes and study success (e.g., invested effort, exam grades, subjective well-being, procrastination) are outlined.

Keywords

Self-regulated learning, assessment, generic competences, higher education, training, strategy use, competence modelling

1 Introduction

The higher education context poses various complex tasks, a great deal of autonomy, and different challenges regarding learning on students (e.g., organizing learning processes, monitoring learning activities, setting goals, using cognitive learning strategies), providing both opportunities as well as the necessity for self-regulated learning (e.g., Peverly et al. 2003). Self-regulated learning (SRL) comprises processes in which learners activate cognitions, affects and behavior that contribute to the attainment of self-imposed goals. Thus, self-regulated learners set goals, use effective learning strategies, monitor their learning process, and create a productive learning environment for themselves (Zimmerman and Schunk 2011). The development of competences for SRL has been set as a goal of higher education (e.g., European Commission 2008) and is also a prerequisite for academic success (e.g., Boekaerts 1997; Wirth and Leutner 2008). A large body of research highlights the importance of SRL processes and components for learning outcomes (e.g., Robbins et al. 2004). High achievement of students is not only associated with high cognitive abilities (e.g., high intelligence), but also with high self-efficacy and high competences in SRL (e.g., the goal-directed use of study and learning strategies; Schneider and Preckel 2017).

In this chapter, the PRO-SRL project¹ is presented which focused on modelling, assessing, and fostering competences for SRL in higher education. The overall project goal was to gain further insight into relevant components of SRL for learning processes in higher education, to develop valid assessments and to gather first evidence in the trainability of these constructs, thereby focusing on the following central research questions:

1. Which competences for SRL are especially important for successful learning processes?
2. How can valid assessments be conceptualized to make competences for SRL more accessible?
3. How can specific competences for SRL (e.g., the regulation of one's own motivation) be trained?

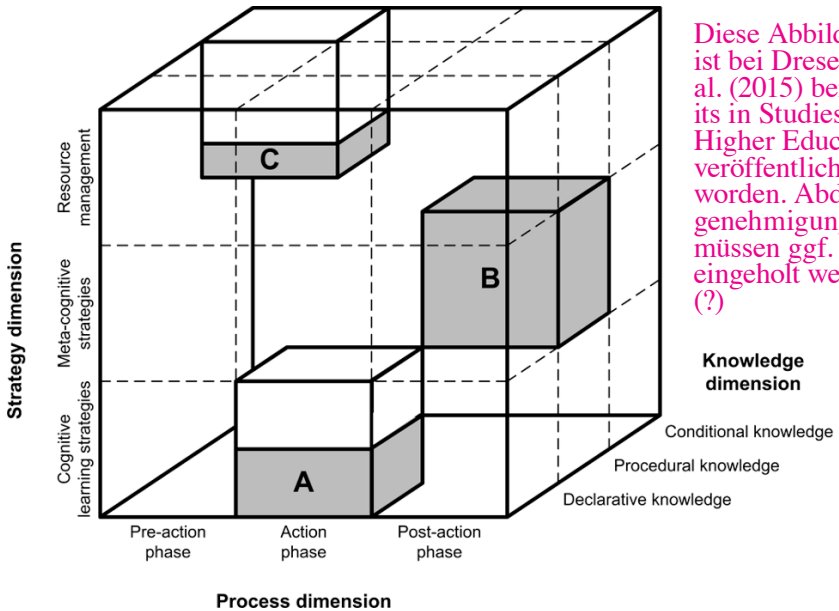
1 The collaborative research project “PRO-SRL – Product- and process oriented modeling and assessment of self-regulation competencies in tertiary education” and its successor “PRO-SRL-EVA – Product- and process oriented modeling and assessment of self-regulation competencies in tertiary education – extended validations” were supported by grants from the German Federal Ministry of Education and Research.

In the following, the results and knowledge gathered throughout the research project regarding these questions will be reported.

2 Competences for SRL

On a behavioral level, the ideal self-regulated learner regulates his or her own learning behavior autonomously, setting adequate learning goals, using learning strategies that are suitable for the learning content and goals, maintaining motivation whilst facing difficulties, and evaluating the learning process and its outcomes afterwards (Artelt et al. 2001). Competences for SRL are considered generic competences that are relevant for all courses of study and should be acquired by every higher education student (Standing Conference of the Ministers of Education and Cultural Affairs 2005). A large body of research attests the importance of these regulatory processes (for an overview, see Zimmerman and Schunk 2011), including evidence for the importance of cognitive regulation (e.g., Schmeck et al. 2014), metacognitive regulation (e.g., Perels et al. 2007) and contextual and personal factors (e.g., dealing with errors, Steuer et al. 2013).

Important models of SRL are process-oriented, focusing on dynamic learning processes with recursive learning cycles that can be divided into different phases with specific regulatory demands: preactional, actional, and postactional (Schmitz and Wiese 2006; Zimmerman 2000). Other models conceptualize SRL by identifying central components: cognitive, metacognitive, and resource-oriented strategies (e.g., Boekaerts 1999; Wild and Schiefele 1994). Since both groups of models have to be considered in combination for a comprehensive understanding of SRL, an integrative model has been developed for the present project (Dresel et al. 2015), uniting process and component models of SRL (Figure 1). The model was proposed as a framework for the development of assessments for different competence facets of SRL. The model also includes different dimensions of strategy knowledge: declarative (knowledge about strategies), procedural (behavioral knowledge about strategies) and conditional strategy knowledge (situation-specific knowledge about when to use a certain strategy; Paris et al. 1983). Thus, competences for SRL are conceptualized as cognitive dispositions to autonomously and appropriately regulate learning processes (Wirth and Leutner 2008) that can be described on three dimensions: process dimension (temporal phases in the cyclic learning process), strategy dimension (the use of cognitive, metacognitive and resource-oriented strategies), and knowledge dimension (declarative, procedural and conditional strategy knowledge).



Diese Abbildung ist bei Dresel et al. (2015) bereits in Studies in Higher Education veröffentlicht worden. Abdruckgenehmigungen müssen ggf. noch eingeholt werden. (?)

Figure 1 Structural model of SRL competences proposed by Dresel et al. (2015), with three marked examples (A, B, and C) for competence facets pronounced to varying extents

In an interview study with 108 experts for different fields of study, Dresel et al. (2015) determined the relevance and suitability of different SRL strategies for different studying situations. The strategies named by the experts fitted to all levels of both the strategy and process dimensions – with experts particularly highlighting the need for resource management strategies (e.g., motivational regulation strategies). Also, the importance of suitability between learning strategies and situational demands was evident. The results were in accordance with the model and confirmed its content validity (Dresel et al. 2015).

3 Conceptualizing SRL Assessments

SRL behavior is often assessed with self-report questionnaires that are characterized by low validity. These questionnaires are frequently suspected to rather capture declarative strategy knowledge (i.e., recognizing strategies) than strate-

gic behavior and are typically marked by low situation specificity (i.e., by asking participants to rate strategy use independent of relevant learning contexts). Nonetheless, these instruments are important for SRL research, as they reveal the intended or remembered strategy use. Moreover, typical questionnaires only address the quantity of strategy use instead of additionally referring to qualitative aspects (Wirth and Leutner 2008). Assessing SRL is especially difficult when the goal is to assess underlying competences (Winne and Perry 2000; Wirth and Leutner 2008). As SRL in general is a cyclical process (e.g., Winne and Hadwin 2008), it is argued for a situation-specific measurement of SRL competences that also considers the different phases of the learning process – making a strong argument for the situational appropriateness of strategy use (Schmitz and Wiese 2006; Wirth and Leutner 2008). Consequently, innovative instruments to assess different aspects of SRL situation-specifically have been developed and validated. After conceptualizing the theoretical basis, several studies were conducted as groundworks for the development of these instruments.

Engelschalk et al. (2015) investigated the importance of situation specificity with regard to motivational regulation. In a self-report study with 54 undergraduates from the field of educational studies, it was examined how students regulate their motivation in different situations (low expectation for success or low subjective task value in three phases of the learning process). Students were asked to report their subjective effectiveness of motivational regulation in the given situations and to describe the strategies they would use. Results revealed that the effectiveness of motivational regulation depended on the specific characteristics of the situation and that students chose different strategies in different situations, which indicates that SRL behavior is perceived situation-specifically by students. For further investigation, Engelschalk et al. (2016) conducted another study with 283 undergraduates from the field of educational studies that also confirmed that motivational regulation is situation-specific. Confirmatory factor analyses showed that the self-reported effectiveness of motivational regulation can be separated for the six different types of motivational problems described above. Results also showed that motivational regulation is perceived as especially difficult when subjective task value is low in the preactional phase (i.e., before a learning activity is started).

Eckerlein et al. (under revision) further investigated the importance of conditional motivational strategy knowledge. Based on the reported findings on the situation specificity of motivational regulation, it was assumed that the suitability of strategies depends on how well they fit to the motivational problem. An expert survey with 33 proven experts in the field of SRL was conducted to determine which strategies are suitable for the proposed six motivational problem situations. The suitability of strategies was indeed considered by the experts to be dependent on

the given motivational problem (e.g., enhancement of personal significance is suitable to increase motivation for a task with low subjective task value). The findings indicate the importance of conditional knowledge in the choice of strategies and further highlight the importance of situation-specific assessment of motivational regulation competences.

Although students know about SRL strategies, they often do not apply them adequately in relevant learning situations. To investigate this discrepancy between knowledge and action and its possible reasons, Foerst et al. (2017) conducted a study with 408 students in two study domains (psychology and economic sciences). Qualitative content analyses confirmed that students had rather advanced knowledge about SRL strategies, although they did not use them extensively in relevant learning contexts. Reported reasons for these discrepancies are the lack of belief in the benefits of using the strategies, no time to use them, not believing in being able to use them effectively, or that their use was too demanding. These findings indicate that assessments for SRL competences should focus on these discrepancies between competence and performance, as they could reveal valuable information on leverage points for fostering SRL, helping higher education students to put their SRL knowledge to action.

4 Assessing Competences for SRL

At the start of the PRO-SRL project, a systematic review of self-report instruments that assess SRL in higher education was conducted (Roth et al. 2015). The results yielded an increasing use of domain-specific measures, but emphasized a lack of scales for assessing motivational and emotional regulation. The majority of the instruments to assess SRL are questionnaires and only very few instruments allow the assessment of SRL closer to actual behavior (e.g., learning journals, interviews, think-aloud techniques). To close these gaps, process-oriented and situation-specific assessments for different SRL facets were developed and validated in the PRO-SLR project.

To assess the differences between knowledge about SRL strategies and their actual use, a competence-performance-assessment (SRL-QuAK) was developed. The instrument allows a simultaneous assessment of competence and performance aspects, based on expert ratings of appropriateness (Foerst et al. 2017). The situation-specific instrument focuses on four aspects of SRL: metacognition, cognition, frustration and boredom. The instrument contains situational vignettes, describing the learning situation and its characteristics (e.g., writing a master thesis). Afterwards, students rate the suitability of certain strategies for the given situation and

also indicate if they are using them. Also, information about why certain strategies are not used is captured, giving relevant information for consulting and training. The internal consistency of the scales was good to excellent (*Cronbach's* $\alpha = .75$ -.94) and first results (e.g., Foerst et al. 2017) showed a high relevance of the differentiation between strategy knowledge and performance. Consistent connections between the scales and achievement ($r = .24$) as well as connections to other relevant concepts (e.g., self-concept of ability, interest, intrinsic motivation) were found for the performance component of SRL strategies.

To assess conditional knowledge regarding motivational regulation, a situational judgement test (Steuer et al. 2019) was developed. The test consists of eight (short version: five) standardized situational vignettes comprising motivational problems while writing a term paper or studying for an exam. Within these situations, different motivational problems are described in different phases of the learning process (expectancy vs. value problems in the preactional vs. actional phase of learning). For each situation description, different strategies for motivational regulation have to be rated regarding their suitability for the given situation. Based on the comparison with expert standards, a score is calculated which reflects the conditional motivational regulation strategy knowledge of the participants. The internal consistency of the test was excellent (*Cronbach's* $\alpha \geq .92$). In a sample of 188 undergraduates in the field of educational studies the strategy knowledge indicator showed positive correlations with motivational regulation on the behavioral level. Moreover, it was positively related to the effectiveness of motivational regulation and effort (Steuer et al. 2019). Convergent connections to other variables could be determined in different studies: quantity of motivational regulation ($r = .46$), effectiveness of motivational regulation ($r = .36$ -.51) and invested effort ($r = .50$; Steuer et al. 2019). In a study by Baulke et al. (2018) conditional motivational regulation strategy knowledge showed significant correlations with procrastination ($r = -.46$) and study dropout intentions ($r = -.23$). Hence, conditional motivational regulation strategy knowledge is relevant and contributes to an extended understanding of motivational regulation in the higher education context.

Apart from conditional knowledge, also the quality of strategy use is important. Regulation quality is conceptualized as planning, implementing, monitoring and, if necessary, adapting strategy use. It is assumed that only coordinated, controlled and accurate strategy use can lead to effective self-regulation (e.g., Pintrich 2005; Winne and Hadwin 2008; Zimmerman 2000). Engelschalk et al. (2017) developed an assessment to capture the quality of motivational regulation strategy use. In a study, 188 students were presented with different situation descriptions posing motivational problems similar to the situations described above. Afterwards they were asked to report both the quantity and quality of motivational regulation strat-

egies they would use in the given situation. The analyses showed that the quantity of strategy use moderately predicted self-reported regulatory effectiveness and study effort. These variables were predicted significantly better by also including the quality of strategy use, which also solely correlated with achievement (Engelschalk et al. 2017).

Further hints for the importance of regulation quality could be derived from a study by Eckerlein et al. (2019). 115 undergraduates from the fields of psychology and educational studies filled in a learning journal during an exam preparation period. Results revealed positive effects of both quantity and quality of motivational regulation on invested effort in exam preparation and on exam performance. Moreover, a high quality of motivational regulation was able to buffer against negative effects of motivational difficulties on invested effort during the learning process.

To capture SRL in process, a standardized learning journal was developed, which allows for fine-grained assessment of SRL over longer periods of time (e.g., during an exam preparation period). The learning journal mainly focusses on metacognitive control (20 items per day). To implement the assessment, it was implemented as a smartphone app (Lang et al. 2017). A strength of this approach is that the app can also be used as a training tool, helping to monitor SRL and giving graphical feedback on SRL behavior.

5 Training of SRL Competences

The PRO-SRL-EVA project also focused on developing training concepts to foster different aspects of self-regulatory competence, which also functioned as validation measures for the developed assessments – in terms of known-group comparisons and testing their change sensitivity. This validation approach also allows us to derive implications for the trainability, the specificity, and the functionality of SRL competences in higher education. The trainings were developed on the basis of the above described competence model and existing literature on fostering SRL (for meta-analyses, see Dignath and Büttner 2008; Dignath et al. 2008).

Although considerable evidence now exists regarding the effects of motivational regulation, little is known about its training. A training study by Eckerlein et al. (submitted) examined whether university students' motivational regulation is trainable and can be improved by means of an approach that addresses three key aspects: the quantity of motivational regulation strategy use, the situation-specific fit between regulation strategy and motivational problems, and the quality of strategy application. A quasi-experimental study with 135 students from the field of educational studies, two experimental conditions (training and placebo group)

and three measuring occasions (pretest, posttest, follow-up) was conducted. The results indicated that the implemented approach was able to improve the quantity and the situation-specific fit of motivational regulation strategy use immediately and sustainably over a six-week period (application quality did not improve significantly). Based on this first training approach, another quasi-experimental study with 131 students from the field of educational studies, two experimental conditions (training, control) and three measuring occasions (pretest, posttest, follow-up) was conducted. The training approach was able to improve the quality and the situation-specific fit of motivational regulation strategy use immediately and over a six-week period (quantity of motivational regulation did not improve significantly). Furthermore, the training had positive and sustainable effects on cognitive and metacognitive strategy use, invested effort, subjective well-being and reduced procrastination. The two training studies indicated that motivational regulation competences are trainable and can be interpreted as further evidence for the validity of the developed instruments to assess situation-specific fit and quality of motivational regulation with which the training effects were identified.

With regard to fostering competences for SRL, several studies are arguing for app-interventions (e.g., Zydney and Warner 2016). As there are no thoroughly evaluated apps for fostering SRL in higher education, Foerst et al. (2019) introduced a smartphone-app that aimed to support students during the process of SRL while writing their bachelor theses. The intervention focused on the promotion of metacognitive SRL-strategies as well as motivational competences. The app consisted of different modules (e.g., planning, setting goals, self-consequation, learning strategies) and was designed to give students the opportunity to constantly monitor and adapt their writing process over the course of one semester. The app was also designed to give graphical feedback on efficiency and strategy use. In a quasi-experimental two-group design (intervention, control) with 118 participants from the fields of psychology and economic sciences who were currently working on their bachelor theses, students received an intervention and used the developed smartphone app. The results indicated a decline in self-reported knowledge about metacognitive strategies and no significant effects for strategy use were found. While motivation rose in the intervention group, unfavorable attribution styles for success and failure were found in both experimental groups. The results give valuable insights into the optimization of smartphone-based interventions regarding the usability and design, which are crucial factors for the use of app-based interventions. One of the key findings was that apps on their own are probably not optimally suited to promote a complex and often almost aversive process, such as SRL. The systematic integration and implementation of the app as a tool in regular teaching seems to be a promising next step in program development.

One important but sometimes neglected component of self-regulation is emotion (Schmitz and Wiese 2006). Well-being can be understood as an overall emotional state relevant for learning (Hascher 2004). As studies show, well-being could enhance other components of SRL and also achievement (Bücker et al. 2018). The aim of a study by Ahrens et al. (in preparation) was to help students to achieve satisfaction and well-being. Earlier research showed that well-being can be enhanced by various interventions in different contexts (Schmitz et al. 2017). For this reason, a diary-based application for smartphones was developed by which well-being should be increased through self-monitoring, well-being exercises and graphical feedback. The exercises trained four components relevant for well-being: meaning, savoring, positive attitude towards life and coping. Through the use of the diary, a daily updated well-being score could be calculated and communicated to the participants in the form of graphical feedback. The effectiveness of the diary-based application to increase well-being was tested with 99 participants who were randomly assigned to three test conditions: (1) self-monitoring, (2) self-monitoring and well-being exercises and (3) self-monitoring, well-being exercises, and graphical feedback. The results of the analyses showed that with the use of the app, well-being and flourishing increased significantly, regardless of the specific condition. The study indicates that well-being can be enhanced in everyday learning situations.

Another promising intervention design is an electronic portfolio which assesses self-reported learning strategy use and helps students to monitor and evaluate their learning behavior (Händel et al. 2018). The portfolio was composed of different parts: a journal (to write about study techniques and strategies), a to-do-list (to set goals and monitor learning) and a discussion forum (to discuss content with peers and ask questions). 1469 students from the field of educational studies used the e-portfolio over nine weeks and the effects of the e-portfolio use on exam grades were analyzed. It was shown that the reported use of cognitive strategies predicted the exam grade, even when controlling for effects of continuous use and time spent working with the e-portfolio. The students using the e-portfolio also showed significantly better achievement in comparison to a control group. The results indicate that online tools that encourage self-monitoring and planning can play an important role in supporting study success and achievement.

6 Summary and Conclusion

Competences for SRL are important prerequisites for the acquisition of subject-specific competences, study success, and life-long learning, as well as an explicit goal of higher education. Research has shown that students substantially differ with regard to SRL and its facets. Grounded on substantial theoretical and empirical work, an integrative theoretical model of SRL competences in higher education, several innovative measuring instruments, as well as training approaches were developed and evaluated in the PRO-SRL and PRO-SRL-EVA projects. The development and extended validation of three measuring instruments (situational judgement test, competence-performance assessment, learning journal) can be considered genuine contributions to the international state of research and useful tools for higher education. The instruments can be used to assess different aspects of self-regulation competences as well as the development of such competences, which is essential to monitor the effectiveness of higher education in attaining the goal of developing SRL competences. Moreover, the instruments could be implemented into higher education as an improved basis of decision-making for fostering SRL, giving higher education personal reliable and valuable hints on the design of learning environments and the consulting of students. Afterwards, the developed assessments can also be used for the evaluation of the effectiveness of implemented training procedures.

To validate the developed assessments, training concepts were designed to foster SRL (e.g., motivational regulation). The trainings had a positive effect on motivational regulation competences, as well as on other aspects of SRL (cognitive and metacognitive strategy use, invested effort, well-being). The developed smartphone-apps pose a mobile approach to fostering not only knowledge but also action components of SRL. The training approaches were developed to be implemented by lecturers and heads of courses of studies. The approaches could be further developed into a bundle of modular training units, which can be chosen based on individual needs of students or institutional standards.

In the future, the instruments and fostering approaches developed in the project could be extended and combined to create a modular training system to foster SRL. Thus, elements could be chosen depending on individual and institutional needs that are identified with the developed assessments (e.g., using a module on setting goals in a class, implementing several modules in a SRL course for undergraduates), allowing for a comprehensive yet fine-grained assessment and training of SRL in higher education.

References

- Ahrens, K., Schäfer, F., & Schmitz, B. (2019). App-based interventions to increase well-being through self-monitoring, well-being exercises and graphical feedback. Manuscript in preparation.
- Artelt, C., Demmrich, A., & Baumert, J. (2001). Selbstreguliertes Lernen. In Deutsches PISA-Konsortium (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* [PISA 2000. Basic competences of pupils in international comparison] (pp. 271–298). Opladen: Leske + Budrich.
- Bäulke, L., Eckerlein, N., & Dresel, M. (2018). Interrelations between motivational regulation, procrastination and college dropout intentions. *Unterrichtswissenschaft*, *46*, pp. 461–479. doi: 10.1007/s42010-018-0029-5
- Boekaerts, M. (1999). Self-regulated learning: Where we are today? *International Journal of Educational Research*, *31*, pp. 445–457. doi: 10.1016/s0883-0355(99)00014-2
- Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*, pp. 161–186. doi: 10.1016/S0959-4752(96)00015-1
- Bücker, S., Nuraydin, S., Simonsmeier, B. A., Schneider, M., & Luhmann, M. (2018). Subjective well-being and academic achievement: A meta-analysis. *Journal of Research in Personality*, *74*, pp. 83–94. doi: 10.1016/j.jrp.2018.02.007
- Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, *3*, pp. 231–264. doi: 10.1007/s11409-008-9029-x
- Dignath, C., Büttner, G., & Langfeld, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on self-regulation training programmes. *Educational Research Review*, *3*, pp. 101–129. doi: 10.1016/j.edurev.2008.02.003
- Dresel, M., Schmitz, B., Schober, B., Spiel, S., Ziegler, A., Engelschalk, T., Jöstl, G., Klug, J., Roth, A., Wimmer, B., & Steuer, G. (2015). Competences for successful self-regulated learning in higher education: Structural model and indications drawn from expert interviews. *Studies in Higher Education*, *40*, pp. 454–470. doi: 10.1080/03075079.2015.1004236
- Eckerlein, N., Engelschalk, T., Steuer, G., & Dresel, M. (2019). Suitability of motivational regulation strategies for specific motivational problems: An expert survey. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*. Manuscript under revision.
- Eckerlein, N., Roth, A., Engelschalk, T., Steuer, G., Schmitz, B., & Dresel, M. (2019). The role of motivational regulation in exam preparation: Results from a standardized journal study. *Frontiers in Educational Psychology*, *10*:8. doi:10.3389/fpsyg.2019.00081
- Eckerlein, N., Steuer, G., & Dresel, M. (2019). Fostering university students' motivational regulation in higher education: Evidence from a quasi-experimental study. *Educational Studies*. Manuscript under revision.
- Engelschalk, T., Steuer, G., & Dresel, M. (2017). Quantity and quality of motivational regulation among university students. *Educational Psychology*, *9*, pp. 1154–1170. doi: 10.1080/01443410.2017.1322177

- Engelschalk, T., Steuer, G., & Dresel, M. (2016). Effectiveness of motivational regulation: Dependence on specific motivational problems. *Learning and Individual Differences*, *52*, pp. 72–78. doi: 10.1016/j.lindif.2016.10.011
- Engelschalk, T., Steuer, G., & Dresel, M. (2015). Wie spezifisch regulieren Studierende ihre Motivation bei unterschiedlichen Anlässen? Ergebnisse einer Interviewstudie. [How specific do students regulate their motivation in different situations? Results from an interview study]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *47*, pp. 14–23. doi: 10.1026/0049–8637/a000120
- European Commission. (2008). *The European Qualifications Framework for Lifelong Learning (EQF)*. Luxembourg: Office for official publications of the European Communities.
- Foerst, N. M., Pfaffel, A., Klug, J., Spiel, C., & Schober, B. (2019). SRL in der Tasche? – Eine SRL-Interventionsstudie im App-Format [SRL in your pocket? An SRL intervention study in app format]. *Unterrichtswissenschaft*, *47*, 337–366. doi:10.1007/s42010–019-00046–7
- Foerst, N. M., Klug, J., Jöstl, G., Spiel, C., & Schober, B. (2017). Knowledge vs. action: Discrepancies in university students' knowledge about and self-reported use of self-regulated learning strategies. *Frontiers in Educational Psychology*, *8*:1288. doi:10.3389/fpsyg.2017.01288.
- Händel, M., Wimmer, B., & Ziegler, A. (2018). E-portfolio use and its effects on exam performance – a field study. *Studies in Higher Education*. Online first publication. doi:10.1080/03075079.2018.1510388
- Hascher, T. (2004). *Wohlbefinden in der Schule* [Well-being in school]. Münster, Deutschland: Waxmann.
- Lang, J., Schumacher, B., & Schmitz, B. (2017, September). *Prozessorientierte Erfassung des Einsatzes selbstregulierter Lernstrategien von Studierenden mithilfe eines als Smartphone-App implementierten Lerntagebuchs* [Process-oriented assessment of students' self-regulation strategy use via learning journal in a smartphone app]. Paper presented at the Conference of the Panels for the German Society for Developmental Psychology and the Panel for Educational Psychology of the German Society of Psychology (DGPS), Münster.
- Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic learner. *Contemporary Educational Psychology*, *8*, pp. 293–316. doi: 10.1016/0361–476X(83)90018–8
- Perels, F., Otto, B., Landmann, M., Hertel, S., & Schmitz, B. (2007). Self-regulation from a process perspective. *Journal of Psychology*, *215*, pp. 194–204. doi: 10.1027/0044–3409.215.3.194
- Peverly, S. T., Brobst K. E., Graham, M., & Shaw, R. (2003). College adults are not good at self-regulation: A study on the relationship of self-regulation, note taking, and test taking. *Journal of Educational Psychology*, *95*, 335–346. doi: 10.1037/0022–0663.95.2.335
- Pintrich, P. R. (2005). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychological and study skill factors predict college outcome? *Psychological Bulletin*, *130*, pp. 261–288. doi:10.1037/0033–2909.130.2.261

- Roth, A., Ogrin, S., & Schmitz, B. (2015). Assessing self-regulated learning in higher education: A systematic literature review of self-report instruments. *Educational Assessment, Evaluation and Accountability*, 3, pp. 225–250. doi: 10.1007/s11092–015-9229–2
- Schmeck, A., Mayer, R. E., Opfermann, M., Pfeiffer, V., & Leutner, D. (2014). Drawing pictures during learning from scientific text: Testing the generative drawing principle and the prognostic drawing principle. *Contemporary Educational Psychology*, 39, pp. 275–286. doi:10.1016/j.cedpsych.2014.07.003
- Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of journal data. *Contemporary Educational Psychology*, 31, pp. 64–96. doi: 10.1016/j.cedpsych.2005.02.002
- Schmitz, B., Lang, J., & Linten, J. (2017). *Psychologie der Lebenskunst: Positive Psychologie eines gelingenden Lebens – Forschungsstand und Praxishinweise* [Psychology of the art of living: Positive psychology for a successful life – State of research and practical implications]. Berlin, Deutschland: Springer.
- Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, 143, pp. 565–600. doi: 10.1037/bul0000098
- Standing Conference of the Ministers of Education and Cultural Affairs (2005). *Qualifikationsrahmen für Deutsche Hochschulabschlüsse [Qualification framework for German university degrees]*. <http://www.kmk.org>
- Steuer, G., Engelschalk, T., Eckerlein, N., & Dresel, M. (2019). Assessment and relationships of conditional motivational regulation strategy knowledge as an aspect of undergraduates' self-regulated learning competences. *Zeitschrift für Pädagogische Psychologie*, 33, pp. 95–104. doi: 10.1024/1010–0652/a000237
- Steuer, G., Rosentritt-Brunn, G., & Dresel, M. (2013). Dealing with errors in mathematics class-rooms: Structure and relevance of perceived error climate. *Contemporary Educational Psychology*, 38, pp. 196–210. doi: 10.1016/j.cedpsych.2013.03.002
- Wild, K.-P., & Schiefele, U. (1994). Lernstrategien im Studium: Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens [Learning strategies in higher education: Results on the factorial structure and reliability of a new questionnaire]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 15, pp. 185–200.
- Winne, P. H., & Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research and applications* (pp. 297–314). New York, NY: Routledge.
- Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531–566). San Diego, CA: Academic Press.
- Wirth, J., & Leutner, D. (2008). Self-regulated learning as a competence: Implications of theoretical models for assessment methods. *Zeitschrift für Psychologie*, 216, pp. 102–110. doi: 10.1027/0044–3409.216.2.102
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego, CA: Academic Press.
- Zimmerman, B., & Schunk, D. (2011). *Handbook of self-regulation of learning and performance*. New York, NY: Routledge.

Zydney, J. M., & Warner, Z. (2016). Mobile apps for science learning: Review of research. *Computers & Education, 94*, pp. 1–17. doi: 10.1016/j.compedu.2015.11.001



3.2

The Relationship between General Intelligence and Media Use among University Students

Jitomirski, J., Zlatkin-Troitschanskaia, O., and Schipolowski, S.

Abstract

Students' information selection process might be influenced by their choice of media sources, their learning contexts and motivation to use certain media as well as their general intelligence, which is crucial for information processing. This study examines the relationship between the general fluid intelligence and the media use of 709 first-year business & economics students from 44 universities in Germany for two different learning purposes: informing oneself about B&E topics and preparing for lectures and exams. Accordingly, the motivator *information seeking* is divided into *curiosity driven* and *goal driven* information seeking. Three types of media sources were included: common news sources, specialized economics sources and university sources. Results from regression analyses and group comparisons indicate that the frequency of media use correlates with general fluid intelligence for some common news sources and specialized economics sources, for example, tabloids and economic newspapers, even after controlling for several sociodemographic variables including gender, age, and parents' educational background.

Keywords

Media use, general fluid intelligence, general intelligence, cognitive ability, higher education, uses and gratifications theory, knowledge acquisition, business and economics

Acknowledgements and Funding Information

The study was funded by the German Federal Ministry of Education and Research under the funding number 01PK15013A.

We would like to thank the anonymous reviewers for their constructive and valuable feedback.

1 Introduction and Research Objectives

Recent years have ushered in the “post-truth era”¹ across all media. The media landscape has shifted accordingly. Especially online, the exponential growth of information sources requires internet users to select their sources critically (Cimpaglia 2018). Information on any website can be presented as authentic regardless of its actual authenticity, causing uncertainty in internet users whether it is genuine. Properly evaluating information has become more challenging and more necessary at the same time (Wineburg et al. 2018; Zlatkin-Troitschanskaia et al. 2019c). Students need appropriate strategies for selecting relevant and reliable information as part of the broader skill of online reasoning and critical thinking – a significant aim of higher education (Gojkov et al. 2015). However, recent studies reveal significant deficits in higher education students’ ability to critically evaluate online media sources (Münchow et al. 2019; Wineburg et al. 2018; Zlatkin-Troitschanskaia et al. 2019c).

The decision which media source students select to consume may be influenced by both their *general intelligence* as well as their underlying *motivation*. Entertainment, social utility, and information seeking are all motivators that affect one’s overall media use (Go et al. 2016; You, Lee et al. 2013). Students’ *information seeking* in particular can be driven by the need to address both personal and study-related inquiries. Students’ media use and how their information seeking

1 *Post-truth* is defined as “relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief” (Oxford English Dictionary 2019).

differs depending on whether it is purely out of curiosity or with a certain goal in mind is of increasing importance due to its potential for informal and formal learning, especially in higher education (for an overview, see Zlatkin-Troitschanskaia et al. 2020).

Students' selection and use of media sources depends on both content and representation of information. Different media sources require different cognitive abilities to process the information they convey, depending on content, language and other characteristics. This leads to the question how students' use of certain sources differs depending on their general intelligence.

This paper focuses on media use of students in *business and economics* (B&E), as this study subject is especially prevalent in the media. B&E topics in general are strongly represented in both mass and social media and are part of the political and societal discourse.² B&E is also one of the biggest study domains worldwide (OECD 2017). Due to the relevance of B&E topics in the media as well as the high study rate, we assessed beginning students from this domain. For B&E students, it is essential to stay informed about current news as well as to read lecture notes and academic literature to acquire study-related content knowledge.

In higher education (including both formal or informal learning), especially the B&E study domain lends itself to gathering data on this topic due to its popularity among students and prevalence in the news cycle (Maurer et al. 2018). As Maurer and colleagues show (2020), in higher education economics, students use certain media to *inform* themselves about B&E topics in general, as well as to *prepare* for their lectures or exams. Based on a theoretical framework described in Section 2, the study investigates B&E students' media use *to inform themselves about B&E topics* as well as *to prepare for lectures and exams*. We focus on the question *to what extent the use of certain media sources differs depending on students' information seeking* behavior and their general intelligence.

2 State of Research in Higher Education and Research Questions

Research literature is ambiguous on the relationship between students' media use and learning in higher education. The displacement hypothesis (Huston et al. 1999) states that using certain media can replace other activities (Blom et al. 2016; Cain and Gradisar 2010; Poulain et al. 2018). For instance, students' media

2 Trade tariffs, the financial crisis, and sanctions against countries are just a few examples of news topics widely discussed in the media today.

use has often been negatively associated with academic performance due to the time spent on consuming media instead of doing homework or paying attention in class (Jacobsen and Forste 2011; Le Roux and Parry 2017; Walsh et al. 2013). This view disregards the numerous possibilities for informal learning through a variety of media sources, especially from media content that relates to the study domain (Maurer et al. 2018). Journalistic news media (newspapers, TV news, magazines) are a potential learning source from which students could benefit (Dalton and Crosby 2013; Kimmerle et al. 2015). Furthermore, using social media such as social networking sites specifically for academic purposes is positively associated with academic achievement (Marker et al. 2018).

According to the Uses and Gratification approach (U&G), every decision to consume media derives from the expectation that a certain need (e.g., knowledge acquisition) will be fulfilled (Saini and Abraham 2019). Based on the U&G theory, there are multiple motivators to consume media (Katz et al. 1974). Many studies have focused on three main ones: information seeking, social utility, and entertainment (Go et al. 2016; You et al. 2013). In particular, different (formal and informal) learning situations and purposes in higher education studies may require different information seeking approaches. University students' behavior may differ depending on the purpose of why they are consuming media content for formal or informal learning.

When asked to rank formal and informal learning situations by their subjective importance, students ranked "reading something" as the most important informal learning situation, while "preparation for an exam" was ranked as the most important formal learning situation (Jadin et al. 2008). Based on these results, this study focuses on two possible learning purposes and corresponding media use motives: to *inform* themselves about B&E topics in general, which in our study is considered an indicator of *curiosity driven* information seeking and students' media use to *prepare* for their lectures or exams, which we consider *goal driven* information seeking.

The frequency of media use allows for inferences about information seeking behavior with regard to *different media sources* (e.g., Schulmeister 2010). In our study, we investigate a relationship between frequency of use of various different media sources and the two different media use motives: curiosity driven and goal driven information seeking. We considered all main types of information sources that university students commonly use according to recent surveys (Maurer et al. 2020; Medienpädagogischer Forschungsverbund Südwest 2017, 2018; Zawacki-Richter 2015). Media sources were divided into three types: common news sources (e.g., local magazines), specialized economics sources (e.g., economic newspapers), and university sources from the formal learning environment (e.g.,

lecture notes and course scripts) (Maurer et al. 2018; Zawacki-Richter 2015). A total of 14 media sources across the three types cover the range students seek information from *to inform themselves* and/or *to prepare for lectures or exams* over their course of study.

Engaging with any type of media poses cognitive demands (Heidi 2018) and requires an adjusted set of *general intellectual abilities* like fluid intelligence, which includes reasoning (Naglieri and Das 2002). (Online) reasoning has been identified as a key ability for processing information from different media sources (Wineburg et al. 2018; Zlatkin-Troitschanskaia et al. 2019). *Reasoning* is considered a proxy for general fluid intelligence as it is especially associated with academic performance (Deary et al. 2007; Rohde and Thompson 2007; te Nijenhuis et al. 2007).

Although many studies found a positive relationship between media use and general intelligence (Beier and Ackerman 2001; Hambrick et al. 2007; Hambrick et al. 2008), the findings do not cover all facets of this multi-dimensional construct. Intelligence can be divided into the facets of fluid and crystallized intelligence (Cattell 1941). Fluid intelligence is the ability to reason and tackle unfamiliar problems (Harrison et al. 2013). A completely novel problem is theoretically solved by reasoning alone, not utilizing previously learned behavior (algorithms, strategies, knowledge, skills etc.) in the process of problem solving. *General crystallized intelligence* involves drawing from long-term memory for information and skills acquired in the past (Jensen 2002).

A relationship between media use and general crystallized intelligence has already been established (Hambrick et al. 2008), while the same authors suggest that general fluid intelligence might also be important for solving novel problems and acquiring new information. The investment hypothesis (Cattell 1963) emphasizes that learners use their general fluid intelligence to invest into the growth of crystallized knowledge, which ultimately promotes the acquisition of knowledge. Although empirically disputed, researchers found a relationship between general fluid intelligence and learning (Schweizer and Koch 2002). Based on this research, this study focuses on students' selection of media and frequency of use, differentiating between two learning purposes (informing oneself about B&E topics, and preparing for lectures or exams), and a relationship to their general fluid intelligence (as a proxy of reasoning). This leads to the following research questions (RQ):

(1) To what extent does general fluid intelligence contribute to explaining the variance in students' media use when (i) *informing themselves* as well as when (ii) *preparing for lectures and/or exams*? (RQ1)

(2) How does the explained variance differ for these two learning purposes and the 14 sources of media – divided into three media types – considered in this study? (RQ2)

(3) Does fluid intelligence remain a significant predictor when including all controlling variables (e.g., gender, age, parents' educational background etc.) considered in this study? (RQ3)

(4) Do groups with the highest and lowest fluid intelligence scores differ systematically in their use of the 14 media sources? (RQ4)

The control variables were chosen based on previous findings and include trust in media, gender, migration background, parental education level, interest in B&E topics and German GPA equivalent. A correlation between trust in media and media use has been found, especially when considering mainstream vs. alternative media (Tsfati and Ariely 2014; Tsfati and Cappella 2003). Gender was controlled for, as previous research has found that males tend to be more interested in B&E topics than females and that interest (in current events) directly affects media use (Hambrick et al. 2008). Therefore, interest in B&E topics was included as a control variable. Moreover, media use has previously been shown to strongly correlate with migration background, and differs greatly between media sources (Bonfadelli et al. 2007; Trebbe 2009). Steiner (2013) has shown that children of parents with a lower level of education consume more media; for instance, they watch more TV. Parental education was measured by asking about the highest level of education within the participants' immediate family. The German GPA equivalent was included as a control variable as a relationship between school grades and general intelligence has been shown in various studies (Colom and Flores-Mendoza 2007; Roth et al. 2015).

3 Method

3.1 Study Design

The data was collected in the WiWiSET study at two measurement points at 44 universities and universities of applied sciences in Germany (Pant et al. 2016). The first measurement took place during the first semester of a bachelor course in B&E studies. A paper-pencil test was distributed to beginning students during introductory classes (Zlatkin-Troitschanskaia et al. 2019b). Under controlled testing conditions, students' general fluid intelligence was assessed using the figural reasoning scale of the Berlin Test for Assessing Crystallized and Fluid Intelligence (BEFKI 11+) (Schipolowski et al. 2020). Additionally, the students provided socio-

demographic and educational information, including their gender, age, migration background, and prior education.

At the end of the first study year, the same students received a link via email to participate in an online questionnaire comprised of multiple validated scales related to their media use (Zlatkin-Troitschanskaia et al. 2019b) that also cover personal and environmental motives of the students' information seeking behavior. Participants received a monetary incentive to participate in the paper-pencil test and the online questionnaire (€5 and €10, respectively).

3.2 Sample

In the large-scale entrance assessment, the total sample of 7,679 first-year students of B&E was collected at the beginning of the winter term of 2016/17 in Germany. 2,584 of these B&E students participated in the first part of the survey at the beginning of their studies. Of these students, 709 participated in both the BEFKI 11+ test and completed the online questionnaire about their media use after one year of studies. The following analysis is based on this subsample collected from 44 universities in Germany.

The participants are between 17 and 37 years old ($M = 20.14$, $SD = 2.40$). The majority of test takers are between 18 and 20 years old (73%), with a median age of 20 years. 96% of the participants being in their first semester of university at the time of the first measurement. 367 participants are female (52%). The average grade of the university entrance qualification is 2.16, with 1.0 being the best and 4.0 the worst possible admission grade. Nearly 50% of the students' parents have a university degree or higher; 179 (25%) students reported a migration background.

3.3 Instruments

3.3.1 Paper-Pencil Test

The figural reasoning scale of the BEFKI 11+ comprised 16 non-verbal items. Each item consisted of a sequence of geometric shapes whose elements changed according to implicit rules. To solve the items, test takers had to infer these rules and choose the next two shapes in the sequence from a number of given alternatives. The 14-minute reasoning test from BEFKI 11+ was chosen as a measure of general fluid intelligence as it can be assumed to be relatively independent of prior education and experience (Gustafsson 1984) and to minimize language ef-

fects (Ackerman and Beier 2003). Although the non-verbal fluid intelligence scale covers fluid intelligence only partly, it is considered to be a valid indicator and overall a reliable proxy for measuring general fluid intelligence under the given time constraints (Carroll 1994; Kyllonen and Christal 1990). *Cronbach's alpha* for the 16-item reasoning test was .66 (for limitations, see Section 5.2).

The students were asked to provide information on their sociodemographic and educational background. The questionnaire included questions about their gender, age, and degree course. To obtain a proxy for socioeconomic status, questions about parental background, for instance migration background and highest education level achieved by mother or father, were included. The questions about prior education surveyed the place and grade of the obtained German GPA-equivalent school leaving certificate.

3.3.2 Online Questionnaire

The online questionnaire included two media use frequency questions separated by the two learning purposes and corresponding subcategories of information seeking: to inform themselves (curiosity driven information seeking) and to prepare for lectures and/or exams (goal driven information seeking) (Maurer et al. 2020; 2018): One question assessed how participants use media to stay informed about B&E topics, and the second question asked which media the students use to prepare for classes and/or exams in B&E. In both questions, 14 media sources were listed (Section 4, Table 1) including journalistic media, for example, common news sources such as national and regional newspapers and TV news, and also specialized domain-specific media sources such as scientific journals, as well as university materials (Maurer et al. 2018). Social networking sites were not included, as they relate primarily to the social utility motivator (Arteaga Sánchez et al. 2014; Mazman and Usluel 2010; Saini and Abraham 2019; Sendurur et al. 2015).

To evaluate the frequency of media use in both cases, the participants were asked how often they had used different types of media during their most recent term at university. The students rated the frequency of their media use for each media type using a scale from 0 (never) to 5 (multiple times a day), with the option of naming additional sources.

To control for students' trust in media sources, the students were asked how trustworthy they considered the previously mentioned media to be on a scale from 0 (not trustworthy at all) to 5 (fully trustworthy) (Rössler 2011). All students rated the same 14 media sources to ensure comparability in the analyses. The online tool provided space to enter other media sources where applicable.

3.4 Analysis

Descriptive statistics provided first insights into the data, including the sample composition and self-reported frequency of media use. The BEFKI 11+ score and information on media use from the online questionnaire were combined into one data set. Correlations between the frequencies of use of the fourteen different media sources were examined to confirm the distinctiveness of the three media types considered in our study. Correlations of media sources within each of the three types (i) common news sources, (ii) specialized economic sources, and (iii) university sources were expected to be higher than correlations between sources from different media types. Correlation analyses also provided information about the distinctiveness of the individual media sources. High correlations between sources ($>.8$) would imply that the sources are too similar and that their overlap is too big to justify their own category.

Subsequently, regression analyses were carried out to examine the relationships proposed in *RQ 1* & *3*. Simple linear regressions were conducted with the BEFKI 11+ score as the independent variable and the media sources divided into two purposes of media use (informing oneself and preparing for exams) as the dependent variables. For *RQ3*, the regression models were extended to multiple linear regression models to include control variables which may contribute to explaining a possible relationship between the frequency of media use and the BEFKI 11+ score. Although ordered logistic regressions would be more suitable for this data, linear regressions have been considered to be an acceptable first approximation (Long and Freese 2006; Pasta 2009). Due to very low intraclass correlations (ICC) of less than .05 across all media sources, multi-level modelling was not conducted (LeBreton and Senter 2008).

In the next step, the link between media use and general fluid intelligence was compared for the students who scored the highest and lowest on the BEFKI test to answer *RQ4*. The participants were divided into quartiles with higher and lower general fluid intelligence scores. A comparison groups approach was used to consider extreme (highest and lowest) scores of students, in particular when examining potential correlations. Preliminary analyses suggested that adding a comparison group approach as a supplement to the regression analyses related to *RQ1* and *RQ3* could increase the statistical power (Preacher et al. 2005). This approach highlights differences within B&E beginning students (Preacher 2014). Apart from the methodological purpose of *RQ4*, the results have the potential to offer initial insights into specific behavioral patterns and systematic differences between two distinct groups. For instance, the findings could lead to possible implications for learning in higher education through media as an enhancement for the lower-scoring group.

The data was analyzed using Stata/IC 15.1. At first, descriptive statistics were calculated for the entire subsample of 709 participants. Additionally, descriptive statistics were calculated for the two “general fluid intelligence” groups (highest and lowest scoring). In addition to simple linear and multiple regression analyses, t-tests were conducted to investigate the research questions.

4 Results

4.1 General Fluid Intelligence as a Factor in Explaining Students’ Media Use

The BEFKI 11+ scores ranged from a minimum of 0 to a maximum of 16. The sample of 709 beginning B&E students achieved a mean score of 9.27 with a standard deviation of 2.75.

To investigate *RQI (i)*, we calculated simple linear regressions to predict the frequency of use of different media to *inform* oneself about B&E topics. In nine out of the 14 listed media sources, the BEFKI 11+ score significantly predicted media use to *inform* oneself. Those nine media sources included both common news and specialized economics sources.

Most variance was significant in the use of tabloids ($F(1,682) = 14.47, p < .001$) with an adjusted R^2 of .02 (Table 6 in App.). It was found that the BEFKI 11+ score significantly predicted frequency of tabloid use ($\beta = -.144, p < .001$). Other sources in which the BEFKI 11+ score predicted more than 1% were news magazines ($\beta = -.07, p < .001$), science journals ($\beta = -.046, p = .001$), and economic ($\beta = -.054, p = .002$), regional ($\beta = -.051, p = .005$), and national newspapers ($\beta = -.053, p = .006$). All relationships were negative, indicating that a higher BEFKI score resulted in lower frequency of media use across all media (Tables 4–12 in App.).

In five out of 14 media sources, the BEFKI 11+ score did not significantly predict media use to *inform* oneself. Those five sources included university materials (textbooks and course scripts), scientific databases and online encyclopedias as well as TV news.

Simple linear regressions for *RQI (ii)* showed that the BEFKI 11+ score predicted the frequency of use of different media to *prepare for classes and/or exams* in eight out of 14 media sources. The relationship was negative across all 14 media sources. Through the BEFKI 11+ score 1.4% of the variance in the use of TV news was predicted ($F(1,682) = 9.66, p = .001$) with an adjusted R^2 of .01 (Table 18 in App.) while the BEFKI 11+ score significantly predicted the use ($\beta = -.076, p = .002$). The variance in the use of economic newspapers ($F(1,682) = 7.73,$

$p = .006$) was predicted with an adjusted R^2 of .01 ($\beta = -.065, p = .006$) and the variance in weekly newspaper use ($F(1,682) = 7.69, p = .006$) with an adjusted R^2 of .01 ($\beta = -.065, p = .006$) (Tables 14 & 15 in App.). Again, the BEFKI 11+ score did not significantly predict the use of university materials such as textbooks and course scripts (for discussion, see Section 5.1).

4.2 Comparison of Information Seeking Motives

For *RQ2*, roughly 80% of students used course scripts almost on a daily basis or even more frequently during the term to both inform themselves about B&E topics (curiosity driven) and to prepare for classes, lectures, and/or exams (goal driven). The difference was not significant in the t-test. Students ranked tabloids as their least used media source for either type of information seeking.

Students used certain media sources more often when informing themselves (*curiosity driven* information seeking), for instance, national and regional newspapers, news magazines, and TV news (Table 1). Students consumed TV news almost daily when motivated by *curiosity driven* information seeking. That frequency went down to a little over once a week during *preparation for classes and/or exams*. The difference between the frequency of curiosity and goal driven TV consumption of news had a medium effect size (*Cohen's d* = .505). The reverse is seen in the use of textbooks. To *inform themselves* about B&E topics students used textbooks roughly a little over a week. During preparation for their studies the use increased to almost daily with a *Cohen's d* = .590, indicating a medium effect size (Cohen 1988).

In total, all common news sources and specialized media sources including university materials were used more frequently during *preparation for classes and/or exams*. Remarkably, some common news sources i.e., tabloids, economic newspapers and TV programs were also used more frequently by students when preparing for exams. Tabloids, although still used least frequently of all sources, were used more often during *preparation for classes and/or exams*. There was no significant difference in the use of video platforms between the information seeking motives.

Correlations between the media sources that students used to *inform themselves* ranged mostly between .2 and .5. An exception was the use of course scripts, which only notably correlated with the use of textbooks (Table 2 in App.). Overall, the correlation analyses confirm the expected distinction of three types among the 14 media sources.

Correlations between the media sources students used to *prepare for exams* were higher across most of them. The correlations between course scripts and

other media stood out, as most were negative (Table 3 in App.). Science data bases proved to be an outlier, correlating relatively highly with all sources except course scripts. Still, the three media types were evident in the correlations between the media sources.

Table 1 Median, mean, and standard deviation for frequency of media use including t-test between media use depending on the information seeking motives

	Media use when informing oneself			Media use when preparing for exams			t(707)	p	d
	Mdn	M	SD	Mdn	M	SD			
National Newspapers	1	2.68	1.42	1	2.36	1.74	4.45	.001	.237
Regional Newspapers	1	2.36	1.31	1	2.25	1.75	1.57	.116	.078
Tabloids	0	1.68	1.16	0	2.04	1.74	-5.22	.001	.277
Economic Newspapers	1	2.26	1.24	1	2.35	1.7	-1.48	.14	.073
Weekly Newspapers	1	2.23	1.24	1	2.3	1.71	-1.02	.308	.045
News Magazines	1	2.65	1.45	1	2.43	1.74	2.87	.004	.152
TV News	3	3.48	1.4	3	2.8	1.8	9.51	.001	.505
Economic TV Programs	1	1.97	1.15	1	2.22	1.71	-3.71	.001	.197
Video Platforms	3	3.42	1.62	3	3.45	1.71	-0.4	.688	.012
Science Journals	0	1.72	1.0	0	2.22	1.74	-7.93	.001	.421
Textbooks	2	2.81	1.34	2	3.47	1.72	-11.09	.001	.59
Course Scripts	3	4.32	1.21	3	4.4	1.65	-1.22	.224	.057
Science Data Bases	0	1.89	1.25	0	2.41	1.83	-8.1	.001	.43
Online Encyclopedia	2	3.33	1.36	2	3.55	1.62	-3.87	.001	.206
Others	0	1.85	1.53	0	2.07	1.75	-3.01	.003	.346

4.3 Additional Influence Factors of Students' Media Use

To investigate *RQ3*, we calculated multiple linear regression models and included control variables to predict the frequency of use of different media to *inform* oneself about B&E topics (Tables 4–11 in App.) as well as to *prepare* for exams (Tables 12–20 in App.). Control variables (trust in media sources, gender, migration background, parents' highest educational level, interest in economics and German GPA equivalent) were included in all regression analyses.

Seven out of 14 media sources to *inform oneself* were significantly explained by the BEFKI 11+ score when the control variables were included in the mod-

els. Two regression models explained more than 15% of the variance: tabloids ($F(7,676) = 19.22, p < .001$) with an adjusted R^2 of .16 ($\beta = -.033, p = .032$) and economic newspapers ($F(7,676) = 17.59, p < .001$) with an adjusted R^2 of .15 ($\beta = -.049, p = .003$) (Tables 6 & 7 in App.), indicating a small effect size (Ellis, 2010). The adjusted R^2 for tabloids was .005 and for economic newspapers .001. The highest adjusted R^2 was .016, for news magazines. The BEFKI 11+ score explained less than 10% of the frequency of media use of all other media sources but remained significant with the exception of video platforms ($F(7,676) = 16.34, p < .001$) with an adjusted R^2 of .14 ($\beta = -.042, p = .051$) (Table 10 in App.). The relationship between the BEFKI 11+ score and the frequency of media use was negative, i.e., the lower the students' intelligence test score was, the more frequently they used media to inform themselves.

Regression analyses predicting the frequency of use of different media to *prepare for classes and/or exams* revealed that the use of six media sources could significantly be predicted by the BEFKI 11+ score but overall only five multiple regression models were significant. Three out of five sources were common news sources: Regional newspapers, news magazines and TV news. The other two sources were economic newspapers and economic TV programs. The highest variance was explained by the model with TV news as the dependent variable ($F(7,676) = 5.04, p < .001$), with an adjusted R^2 of .04 ($\beta = -.079, p = .002$) (Table 17 in App.). This media source also showed the highest adjusted R^2 (.012). In total significant variance explained ranged from 1.2 to 4%. The BEFKI 11+ score did not provide any significant predictions for any media sources from the university formal learning environment, the use of national and weekly newspapers, or video platforms when all previously mentioned control variables were included.

Additional factors that significantly predict how students *inform themselves* are their *general interest in B&E topics* and *trust* in the media source in question. Interest in B&E is especially predictive for specialized economic sources, for instance, economic newspapers ($\beta = .257, p < .001, \text{Adj. } R^2 = .06$), economic TV programs ($\beta = .16, p < .001, \text{Adj. } R^2 = .02$) and science journals ($\beta = .191, p < .001, \text{Adj. } R^2 = .03$). Trust in a media source predicts the use of material that typically has fewer built-in filter systems i.e., tabloids ($\beta = .35, p < .001, \text{Adj. } R^2 = .13$), regional newspapers ($\beta = .174, p < .001, \text{Adj. } R^2 = .03$) and video platforms ($\beta = .255, p < .001, \text{Adj. } R^2 = .06$) (Tables 4 – 11 in App.).

Trust also predicted media use in *preparation for lectures or exams* for different media sources including news magazines ($\beta = .09, p = .019, \text{Adj. } R^2 = .01$), TV news ($\beta = .175, p < .001, \text{Adj. } R^2 = .3$), economic TV programs ($\beta = .082, p = .032, \text{Adj. } R^2 = .01$) and video platforms ($\beta = .278, p < .001, \text{Adj. } R^2 = .08$). In contrast to media use to inform oneself, parental education played an important role in

predicting the frequency of media use when preparing for lectures and exams. The frequency of use for regional newspapers ($\beta = -.085, p = .03, \text{Adj. } R^2 = .01$), economic TV programs ($\beta = -.096, p = .014, \text{Adj. } R^2 = .01$), and science journals ($\beta = -.1, p = .01, \text{Adj. } R^2 = .01$) was higher where the parents' level of education was lower (Tables 12 – 20 in App.).

4.4 Comparison of General Fluid Intelligence Groups and Their Media Use

To investigate *RQ4*, the subsample was divided into four groups by quartiles depending on their BEFKI 11+ score. Group 1 included the highest scoring quarter of participants ($n_1 = 160, M = 12.9, SD = 1.02$) with scores ranging from 12 to 16. Group 2 consisted of the lowest scoring quarter of participants with scores ranging from 0 to 7 ($n_2 = 185, M = 5.79, SD = 1.38$). The two groups did not significantly differ in terms of age, gender, or attending higher-level economics courses in high school. However, they differed in their obtained German GPA equivalent. Group 1 had a better GPA ($M = 1.95, SD = .56$) than group 2 ($M = 2.32, SD = .56$), $t(339) = 6.1357, p < .001$. The two groups also differed in the number of students with a migration background. While 18.76% of the students in group 1 reported that they have at least one parent that was not born in Germany, more than one third of the students in group 2 reported that they have a migration background (36.76%). 21.08% of the students in group 2 had completed vocational training in B&E prior to starting their university course, while only 9.38% group 1 had done so. T-tests were calculated on the basis of the comparison groups approach to explore whether the high general fluid intelligence group (group 1) and the low general fluid intelligence group (group 2) differed systematically. Here, first insights on specific behavioral patterns could be gained in addition to the results of regression analyses related to *RQ1* and *RQ3*. Students in the upper quartile of the BEFKI 11+ score consistently used all 14 media sources less frequently than students in the lower quartile. This applies to both learning purposes, when informing oneself and preparing for exams.

The groups differed significantly in their use of most media sources to inform oneself, except for TV news, course scripts, and online encyclopedias (Table 21 in App.). The highest significant difference was found in the use of tabloids. These findings are consistent with the results related to *RQ1* and *RQ3*. Group 2 used tabloids more ($M = .98, SD = 1.42$) than group 1 ($M = .39, SD = .10$), $t(343) = 4.49, p < .001, \text{Cohen's } d = .485$, indicating a medium effect size (Cohen, 1988). The two groups differ significantly in their media use of textbooks. Group 1 used textbooks

out of a general interest to be *informed* about B&E topics less often ($M = .53$, $SD = .76$) than group 2 ($M = .89$, $SD = 1.09$), $t(343) = 3.56$, $p < .001$, *Cohen's* $d = .385$, indicating a small effect size (Cohen, 1988), although it is on the slightly higher end of the interpretation of effect size. Due to the increased statistical power of the comparison groups approach, the use of university sources can be further differentiated between the groups.

Analyses of the groups concerning their media use to prepare for exam and/or lectures revealed statistically significant differences between the two groups in only five of the 14 media sources: Regional, economic, and weekly newspapers, news magazines, and TV news (Table 22 in App.). None of these were university sources and only one was a specialized economics source. The greatest difference between the groups was found in the use of TV news. Group 2 used TV news more frequently ($M = 2.13$, $SD = 1.82$) than group 1 ($M = 1.62$, $SD = 1.84$), $t(343) = 2.59$, $p = .010$, *Cohen's* $d = .28$ indicating a small effect size (Cohen 1988). The other four media sources (regional, economic, and weekly newspapers, news magazines) had effect sizes ranging from .228 (news magazines) to .266 (weekly newspapers). These findings are also consistent with the results related to *RQ1* and *RQ3*, even though the regression analyses related to *RQ1* and *RQ3* suggest that the BEFKI 11+ score can predict the use of more media sources for the purpose of preparing for exams and/or lectures.

Overall, the differences in the frequency of media use to inform oneself were more substantial and showed higher effect sizes in 12 of the 14 media sources. In comparison, the groups only differed in five out of 14 media sources significantly concerning the preparation for lectures and/or exams with small effect sizes.

5 Discussion

5.1 Media Use and General Fluid Intelligence of Students

Regarding *RQ1(i)*, the relationship between students' media use to *inform* themselves about B&E topics (*curiosity driven information seeking*) and general fluid intelligence differs depending on the media source, although the relationship was always negative. Nine out of 14 media sources had a significant relationship with general fluid intelligence, including common news and specialized economics sources, while university sources were not significantly predicted by the BEFKI 11+ score. Overall, general fluid intelligence appears to predict media use to inform oneself if the media source is less directly related to formal university learning. General fluid intelligence seems to influence the use of media sources if

students want to inform themselves, as is evidenced by the fact that students with higher scores have a lower frequency of media use. This is especially the case for media sources for which the lowest content quality is expected – tabloids – which supports the general assumption that certain media pose different cognitive demands. Students with lower cognitive abilities seem to prefer consuming media that is less demanding.

There is a correlation between general fluid intelligence and the use of regional, economic, and weekly newspapers as media sources when preparing for lectures and exams (*RQ1(ii)*). Similar to the findings related to *RQ1(i)*, the correlation was negative, i.e., the lower the students' BEFKI 11+ score was, the more frequently they used common media sources (e.g., TV news). Since these media can be expected to be of little help for exam preparation, one interpretation of this finding could be that a more "intelligent" strategy would be to use these media sources less and to concentrate more on media sources that contain more study-related and domain-specific content. However, general fluid intelligence did not significantly predict the use of university materials. This result suggests that students with different levels of general fluid intelligence use the same sources to study for university.

The findings of the present study support the distinction between the U&G information seeking motives and the assumption that they influence the students' media use behavior (*RQ2*). In general, the media use across most media sources was more frequent when the process of seeking information was driven by a study-specific goal. A few common news sources were an exception: national and regional newspapers, news magazines, and TV news, which were used more often by students to inform themselves in general. These sources are seemingly not considered by students to be viable options for preparing for lectures and exams, which suggests that they are less likely to use these common news sources to supplement their formal learning.

Generally speaking, general fluid intelligence can account for slightly more variance in media use when informing oneself than when preparing for lectures and exams. One possible explanation for this difference might be the selection process of reliable media sources for curiosity driven information seeking. Students with a higher level of fluid intelligence seem to have identified their trusted media source. Also, they seem to be less likely to "shop around" for the right media source, as a critical selection of their favored media source has already taken place, leading to an overall lower frequency of media use. Findings also suggest that students with a higher level of fluid intelligence consume less media overall, independent of their information seeking motive or media source. One reason for this might be that this group of students might select media sources and acquire and process new information more effectively or faster.

To answer *RQ3*, multiple regressions that included control variables were conducted. Approximately 15 % of the variance of media use when informing oneself was explained for tabloids, economic newspapers, and science journals. The adjusted R^2 for tabloids was .005, for economic newspapers .01, and for science journals .01. There was no significant relationship between the BEFKI+ score and the use of university materials such as textbooks and course scripts even when adding control variables. There was, however, a significant negative relationship between using economic newspapers and general fluid intelligence. Students with a lower level of general fluid intelligence might prefer economic newspapers to university media sources due to their comparatively simple representations of B&E topics. Study materials do not cover current news topics and usually convey basic principles and theories so students can acquaint themselves with the subject domain, which might explain the overall lower frequency of reported media use. Our data, however, do not include information about the quality of media use i.e. how attentively a student might read the text and about the texts students have actually consumed, only about the frequency of use and the type of media sources (for the limitations, see Section 5.2). For instance, students with a higher level of general fluid intelligence might read economic newspapers more rarely but more thoroughly; students with a lower level of general fluid intelligence might need more time to process novel information. This study does not allow for more in-depth conclusions about students' information processing or for explanations regarding the negative relationship between general fluid intelligence and media use for informing oneself about B&E topics.

Students who have more difficulties processing novel information might prefer non-university materials due to simpler or more user-friendly representations of information. Using tabloids is negatively associated with general fluid intelligence, as this medium may provide more easily understandable information regarding B&E and is often written in a flowery, not very fact-based journalistic style. Staying informed (using non-university sources) and keeping up with current trends using tabloids might be connected to the displacement hypothesis (Huston et al. 1999). The use of tabloids as a media source may also be more strongly connected to a different motivator, for example, entertainment, which was not considered in this study. Although students may feel like they are informing themselves about B&E topics, the content of tabloids is less likely to cover this domain in depth. This stereotype about what this media source can offer, backed by the displacement hypothesis of foregoing one activity for another, might offer one initial explanation.

Adding control variables to explain media use when preparing for lectures and exams led to general fluid intelligence not significantly predicting some of the media sources, for example, national and weekly newspapers, video platforms,

and science journals. The remaining significant sources included common news sources and specialized economic sources. The regression model with TV news as dependent variable explained most of the variance. Again, the use of media sources that were not traditionally used for exam preparation negatively correlated with general fluid intelligence. One possible explanation might be a subjective feeling of students with lower fluid intelligence that consuming certain non-university media sources to prepare for lectures and exams is useful. They might rely more on informal or passive learning through media in addition to university sources as a learning strategy, thus spreading their own information processing resources thinner.

Regarding *RQ4*, students in the lower quartile of general fluid intelligence (group 2) consistently reported that they use more media from all 14 sources than the students with a higher level of general fluid intelligence (group 1) when informing themselves. The two groups differed significantly in their use of most media sources except TV news, course scripts, and online encyclopedias. The most significant difference is that students in group 2 used more tabloids and textbooks, although both effect sizes are small. Unlike in *RQ1* and *RQ3*, a difference in the frequency of textbook use emerged for *RQ4*. Although textbooks are a reliable source of information, using them to inform oneself in general might suggest two things. For one, students might consider common news sources, for instance, newspapers or specialized economics sources, a more appropriate choice to obtain current information. Moreover, university sources are more easily available to students and are already “approved” by faculty staff, which might make a selection process of suitable media sources easier, particularly for students in the lower general fluid intelligence group. Students with a low level of general fluid intelligence consistently used all 14 media sources more frequently to prepare for lectures and exams compared to the students with a high level of general fluid intelligence. Statistically significant differences were determined for five of the 14 media sources: regional, economic, and weekly newspapers, news magazines, and TV news. This finding is consistent with the regression analyses. The most notable difference between groups 1 and 2 was identified in the use of TV news, with a small effect size. This finding is in line with previous studies suggesting that groups with a lower educational level process TV news best, while showing a smaller processing capacity for newspapers (Grabe et al. 2009). The groups did not significantly differ in their use of university materials. However, media use differs in sources (e.g., regional, economic, and weekly newspapers, news magazines, and TV news) that might not be specifically needed for achieving a study-related goal such as preparing for a lecture or passing an exam.

5.2 Limitations and Future Directions

The survey data only show a self-reported estimate of students' frequency of media use. Although there is a discrepancy between self-reported measures of media use and actual media use, studies have found strong positive correlations between self-reported and actual time spent consuming different media sources (e.g., Facebook: Junco 2013). Nonetheless, future studies should use different measurements that provide more objective and detailed data, for instance, by asking students to report the concrete duration of use, using a continuous measure or by asking students how many minutes in the last week they spent using certain media, as well as by tracking and analyzing the content and quality of used media. Many of the media surveyed may have content that requires processing at very different (cognitive) levels. New technologies, such as apps like Social Fever and MyAddictometer (Basera 2019), make it possible to track media use in real-time.

Furthermore, traditional media sources have increasingly moved into the online space (Fletcher and Park 2017). The fast-changing media landscape invites active participation rather than passive use. Further research is required on the expansion of the U&G approach presented in this article in combination with a differentiation between active and passive media use. In this study, only two meta-categories of learning purposes were considered, which also require more detailed differentiation and examination in different formal and informal learning situations.

More detailed investigations as to which media from the different categories in particular influenced media use and, for instance, further breaking down the category of news magazines into concrete well-known German magazines like *Der Spiegel*, *Stern*, and *Focus* would provide a more nuanced insight into students' media use and allow for in-depth analyses regarding the potential of informal learning through mass media.

The results suggest that in future studies, general fluid intelligence should be measured more comprehensively and used as a covariate when analyzing media use and its influence on student learning. In this study, only one core facet of general intelligence is measured, using a figural reasoning scale as a proxy. Further studies are required which include both additional scales of figural reasoning as well as measurement of other facets of general intelligence. For instance, reading and processing texts in particular may be more strongly related to verbal intelligence.

In addition, as students' media use was assessed only at one measurement point in this study, the findings do not allow for statements about causality. For instance, since the students in group 1 (high general fluid intelligence group) generally used less media from all 14 listed media sources, future research is required to examine media use in relation to general intelligence over the course of study. A longitu-

dinal study with more measurement points, for example spanning the duration of university studies, would allow for initial causality analyses.

Finally, studies with students from other study domains are required to examine to which extent the findings of this study are domain- and B&E-content specific, and which media use patterns may become evident among students from other study domains in different learning situations (for a comparative analysis with social science students, see Maurer et al. 2020).

5.3 Conclusion

This article aimed to examine the relationships between students' media use and their general fluid intelligence, differentiating between two learning purposes: informing themselves (curiosity driven information seeking) and preparing for lectures or exams (goal driven information seeking). This distinction is related to the Uses and Gratification approach (U&G), which suggests different media use motivation, one of them being information seeking motives. Curiosity driven media use can be linked to learning when informing oneself, while goal driven media use can be related to learning when preparing for exams.

The study has shown that depending on the learning purpose, i.e. what drives students to seek out information, they *(i)* use media differently and *(ii)* use different media. The curiosity driven use of media has a stronger relationship with general fluid intelligence than goal driven media use. The patterns of media use depending on the information seeking motive is the same for higher and lower scoring students on a general fluid intelligence test.

For higher education practice, this study emphasizes the crucial importance of establishing a habit of conscious use of media facilitated by academic staff. This can be done by deliberately integrating mass media sources into the learning context as well as by encouraging informal media use, i.e., informal learning through media sources that are not directly study-related. This includes utilizing mass media sources to support habits of active knowledge transfer from one domain to the other to create synergies between the two. Building the habit of knowledge transfer and a strengthened ability to evaluate media sources as conscious media consumers in different domains could support learning in higher education.

References

- Ackerman, P. L., & Beier, M. E. (2003). Trait complexes, cognitive investment and domain knowledge. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 1–30). Cambridge: Cambridge Univ. Press.
- Arteaga Sánchez, R., Cortijo, V., & Javed, U. (2014). Students' perceptions of Facebook for academic purposes. *Computers & Education*, 70, 138–149. <https://doi.org/10.1016/j.compedu.2013.08.012>
- Basera, B. (2019). *10 Best Apps That Track And Limit Social Media Usage*.
- Beier, M. E., & Ackerman, P. L. (2001). Current-events knowledge in adults: An investigation of age, intelligence, and nonability determinants. *Psychology and Aging*, 16(4), 615–628. <https://doi.org/10.1037/0882-7974.16.4.615>
- Blom, N., van der Zanden, R., Buijzen, M., & Scheepers, P. (2016). Media Exposure and Health in Europe: Mediators and Moderators of Media Systems. *Social Indicators Research*, 126, 1317–1342. <https://doi.org/10.1007/s11205-015-0933-6>
- Bonfadelli, H., Bucher, P., & Piga, A. (2007). Use of old and new media by ethnic minority youth in Europe with a special emphasis on Switzerland. *Communications*, 32(2), 459. <https://doi.org/10.1515/COMMUN.2007.010>
- Cain, N., & Gradisar, M. (2010). Electronic media use and sleep in school-aged children and adolescents: A review. *Sleep Medicine*, 11(8), 735–742. <https://doi.org/10.1016/j.sleep.2010.02.006>
- Carroll, J. B. (1994). Cognitive abilities: Constructing a theory from data. In D. K. Detterman (Eds.), *Current topics in human intelligence ; vol. 4: Theories in intelligence* (pp. 43–63). Norwood, NJ: Ablex.
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1–22. <https://doi.org/10.1037/h0046743>
- Ciampaglia, G. (2018). The Digital Misinformation Pipeline. In O. Zlatkin-Troitschanskaia, G. Wittum, & A. Dengel (Eds.), *Positive Learning in the Age of Information* (pp. 413–422). Wiesbaden: Springer.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.
- Colom, R., & Flores-Mendoza, C. E. (2007). Intelligence predicts scholastic achievement irrespective of SES factors: Evidence from Brazil. *Intelligence*, 35(3), 243–251. <https://doi.org/10.1016/j.intell.2006.07.008>
- Dalton, J. C., & Crosby, P. C. (2013). Digital Identity: How Social Media Are Influencing Student Learning and Development in College. *Journal of College and Character*, 14(1), 1–4. <https://doi.org/10.1515/jcc-2013-0001>
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). *Intelligence and educational achievement*. *Intelligence*, 35(1), 13–21. <https://doi.org/10.1016/j.intell.2006.02.001>
- Ellis, P. D. (2010). *The Essential Guide to Effect Sizes*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511761676>
- Fletcher, R., & Park, S. (2017). The Impact of Trust in the News Media on Online News Consumption and Participation. *Digital Journalism*, 5(10), 1281–1299. <https://doi.org/10.1080/21670811.2017.1279979>
- Go, E., You, K. H., Jung, E., & Shim, H. (2016). Why do we use different types of websites and assign them different levels of credibility? Structural relations among users' motives,

- types of websites, information credibility, and trust in the press. *Computers in Human Behavior*, 54, 231–239. <https://doi.org/10.1016/j.chb.2015.07.046>
- Gojkov, G., Stojanović, A., & Rajić, A. G. (2015). Critical Thinking of Students – Indicator of Quality in Higher Education. *Procedia – Social and Behavioral Sciences*, 191, 591–596. <https://doi.org/10.1016/j.sbspro.2015.04.501>
- Grabe, M. E., Kamhawi, R., & Yegiyani, N. (2009). Informing Citizens: How People with Different Levels of Education Process Television, Newspaper, and Web News. *Journal of Broadcasting & Electronic Media*, 53(1), 90–111. <https://doi.org/10.1080/08838150802643860>
- Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, 8(3), 179–203. [https://doi.org/10.1016/0160-2896\(84\)90008-4](https://doi.org/10.1016/0160-2896(84)90008-4)
- Hambrick, D. Z., Meinz, E. J., & Oswald, F. L. (2007). Individual differences in current events knowledge: Contributions of ability, personality, and interests. *Memory & Cognition*, 35(2), 304–316. <https://doi.org/10.3758/BF03193451>
- Hambrick, D. Z., Pink, J. E., Meinz, E. J., Pettibone, J. C., & Oswald, F. L. (2008). The roles of ability, personality, and interests in acquiring current events knowledge: A longitudinal study. *Intelligence*, 36(3), 261–278. <https://doi.org/10.1016/j.intell.2007.06.004>
- Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science*, 24(12), 2409–2419. <https://doi.org/10.1177/0956797613492984>
- Heidi, J. (2018). Taking a Stand in the Post-Truth Era. *Pakistan Journal of Information Management & Libraries*, 20, i.
- Huston, A. C., Wright, J. C., Marquis, J., & Green, S. B. (1999). How young children spend their time: Television and other activities. *Developmental psychology*, 35(4), 912–925. <https://doi.org/10.1037/0012-1649.35.4.912>
- Jacobsen, W. C., & Forste, R. (2011). The wired generation: Academic and social outcomes of electronic media use among university students. *Cyberpsychology, Behavior and Social Networking*, 14(5), 275–280. <https://doi.org/10.1089/cyber.2010.0135>
- Jadin, T., Richter, C., & Zöserl, E. (2008). Formelle und informelle Lernsituationen aus Sicht. In S. Zauchner (Eds.), *Medien in der Wissenschaft: Vol. 48. Offener Bildungsraum Hochschule: Freiheiten und Notwendigkeiten ; [13. Europäische Jahrestagung der Gesellschaft für Medien in der Wissenschaft (GMW08)]*. Münster [u.a.]: Waxmann.
- Jensen, A. R. (2002). Psychometric g: Definition and Substantiation. In R. J. Sternberg & E. Grigorenko (Eds.), *The general factor of intelligence: How general is it?*. Mahwah, NJ: Erlbaum.
- Junco, R. (2013). Comparing actual and self-reported measures of Facebook use. *Computers in Human Behavior*, 29(3), 626–631. <https://doi.org/10.1016/j.chb.2012.11.007>
- Kimmerle, J., Moskaliuk, J., Oeberst, A., & Cress, U. (2015). Learning and Collective Knowledge Construction With Social Media: A Process-Oriented Perspective. *Educational Psychologist*, 50(2), 120–137. <https://doi.org/10.1080/00461520.2015.1036273>
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!. *Intelligence*, 14(4), 389–433. [https://doi.org/10.1016/S0160-2896\(05\)80012-1](https://doi.org/10.1016/S0160-2896(05)80012-1)

- Le Roux, D. B., & Parry, D. A. (2017). In-lecture media use and academic performance: Does subject area matter?. *Computers in Human Behavior*, 77, 86–94. <https://doi.org/10.1016/j.chb.2017.08.030>
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods*, 11(4), 815–852. <https://doi.org/10.1177/1094428106296642>
- Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2. ed.). College Station, Tex.: Stata Press Publ.
- Marker, C., Gnamb, T., & Appel, M. (2018). Active on Facebook and Failing at School? Meta-Analytic Findings on the Relationship Between Online Social Networking Activities and Academic Achievement. *Educational Psychology Review*, 30(3), 651–677. <https://doi.org/10.1007/s10648-017-9430-6>
- Maurer, M., Quiring, O., & Schemer, C. (2018). Media Effects on Positive and Negative Learning. In O. Zlatkin-Troitschanskaia, G. Wittum, & A. Dengel (Eds.), *Positive Learning in the Age of Information*. Wiesbaden: Springer.
- Maurer, M., Schemer, C., Zlatkin-Troitschanskaia, O., & Jitomirski, J. (2020). Positive and Negative Media Effects on University Students' Learning: Preliminary Findings and a Research Program. In O. Zlatkin-Troitschanskaia, (Ed.), *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*. Springer.
- Mazman, S. G., & Usluel, Y. K. (2010). Modeling educational usage of Facebook. *Computers & Education*, 55(2), 444–453. <https://doi.org/10.1016/j.compedu.2010.02.008>
- Medienpädagogischer Forschungsverbund Südwest (2017). *JIM-Studie 2017 – Jugend, Information, (Multi-)Media.: Basisstudie zum Medienumgang 12- bis 19-Jähriger in Deutschland 2017*. Retrieved from https://www.mpfs.de/fileadmin/files/Studien/JIM/2017/JIM_2017.pdf
- Medienpädagogischer Forschungsverbund Südwest (2018). *JIM-Studie 2018 – Jugend, Information, Medien: Basisuntersuchung zum Medienumgang 12- bis 19-Jähriger*. Retrieved from https://www.mpfs.de/fileadmin/files/Studien/JIM/2018/Studie/JIM_2018_Gesamt.pdf
- Münchow, H., Richter, T., Mühlen, S. von der, & Schmid, S. (2019). The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network, and relevance for academic success at the university. *The British Journal of Educational Psychology*. Advance online publication. <https://doi.org/10.1111/bjep.12298>
- Naglieri, J. A., & Das, J. P. (2002). Practical Implications of General Intelligence and PASS Cognitive Processes. In R. J. Sternberg & E. Grigorenko (Eds.), *The general factor of intelligence: How general is it?*. Mahwah, NJ: Erlbaum.
- OECD. (2017). *Education at a Glance 2017: OECD Indicators*. Paris: OECD Publishing.
- Oxford English Dictionary. (2019). “*post-truth, adj.*”: *Oxford University Press*. Retrieved from <https://www.oed.com/view/Entry/58609044?redirectedFrom=post-truth&>
- Pant, H. A., Zlatkin-Troitschanskaia, O., Schipolowski, S., & Förster, M. (2016). WiWiSET – Validation of an Entrance Examination in the Study Domain of Business and Economics and Universities of Applied Sciences: A National and International Comparative Study of Universities. In H. A. Pant, O. Zlatkin-Troitschanskaia, C. Lautenbach, M. Toepper, & D. Molerov (Eds.), *Modeling and Measuring Competencies in Higher Education – Validation and Methodological Innovations (KoKoHs): Overview of the Research Pro-*

- jects* (pp. 60–61). Retrieved from https://www.kompetenzen-im-hochschulsektor.de/files/2018/05/KoKoHs_Working_Papers_No.11_Final_04.07.pdf
- Pasta, D. J. (2009). *Learning when to be discrete: Continuous vs. categorical predictors. Proceedings of the Thirty-fourth Annual (No. 248)*. Washington, DC, USA.
- Poulain, T., Peschel, T., Vogel, M., Jurkatat, A., & Kiess, W. (2018). Cross-sectional and longitudinal associations of screen time and physical activity with school performance at different types of secondary school. *BMC Public Health*, 18(1), 563. <https://doi.org/10.1186/s12889-018-5489-3>
- Preacher, K. J. (2014). Extreme Groups Designs. In R. L. Cautin (Eds.), *The encyclopedia of clinical psychology* (Vol. 40, pp. 1–4). Chichester: Wiley. <https://doi.org/10.1002/9781118625392.wbecp190>
- Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, 10(2), 178–192. <https://doi.org/10.1037/1082-989X.10.2.178>
- Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, 35(1), 83–92. <https://doi.org/10.1016/j.intell.2006.05.004>
- Rössler, P. (2011). *Skatenhandbuch Kommunikationswissenschaft* (1. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften/Springer Fachmedien Wiesbaden GmbH Wiesbaden. Retrieved from <http://dx.doi.org/10.1007/978-3-531-94179-0> <https://doi.org/10.1007/978-3-531-94179-0>
- Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118–137. <https://doi.org/10.1016/j.intell.2015.09.002>
- Saini, C., & Abraham, J. (2019). Modeling educational usage of social media in pre-service teacher education. *Journal of Computing in Higher Education*, 31(1), 21–55. <https://doi.org/10.1007/s12528-018-9190-4>
- Schipolowski, S., Wilhelm, O., & Schroeders, U. (2020). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 11. und 12. Jahrgangsstufe (BEFKI 11–12+)*. Göttingen: Hogrefe.
- Schulmeister, R. (2010). Students, Internet, eLearning and Web 2.0. In M. Ebner & M. Schiefner (Eds.), *Looking Toward the Future of Technology-Enhanced Education* (pp. 317–347). IGI Global.
- Schweizer, K., & Koch, W. (2002). A revision of Cattell's Investment Theory. *Learning and Individual Differences*, 13(1), 57–82. [https://doi.org/10.1016/S1041-6080\(02\)00062-6](https://doi.org/10.1016/S1041-6080(02)00062-6)
- Sendurur, P., Sendurur, E., & Yilmaz, R. (2015). Examination of the social network sites usage patterns of pre-service teachers. *Computers in Human Behavior*, 51, 188–194. <https://doi.org/10.1016/j.chb.2015.04.052>
- Steiner, O. (2013). Pflegen medienkompetente Eltern eine gute Medienerziehung? Ergebnisse einer Repräsentativbefragung von Eltern 10- bis 17-jähriger Kinder. *Diskurs Kindheits- und Jugendforschung/Discourse*, (8), 471–484.
- Te Nijenhuis, J., van Vianen, A. E.M., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, 35(3), 283–300. <https://doi.org/10.1016/j.intell.2006.07.006>
- Trebbe, J. (2009). *Ethnische Minderheiten, Massenmedien und Integration: Eine Untersuchung zu massenmedialer Repräsentation und Medienwirkungen*. Zugl.: Berlin, Freie Univ., Habilitationsschrift, 2008. Wiesbaden: VS Verlag für Sozialwissenschaften/GWV Fachverlage GmbH Wiesbaden. <https://doi.org/10.1007/978-3-531-91696-5>

- Tsfati, Y., & Ariely, G. (2014). Individual and Contextual Correlates of Trust in Media Across 44 Countries. *Communication Research*, 41(6), 760–782. <https://doi.org/10.1177/0093650213485972>
- Tsfati, Y., & Cappella, J. N. (2003). Do People Watch what they Do Not Trust? *Communication Research*, 30(5), 504–529. <https://doi.org/10.1177/0093650203253371>
- Walsh, J. L., Fielder, R. L., Carey, K. B., & Carey, M. P. (2013). Female College Students' Media Use and Academic Outcomes: Results from a Longitudinal Cohort Study. *Emerging Adulthood (Print)*, 1(3), 219–232. <https://doi.org/10.1177/2167696813479780>
- Wineburg, S., Breakstone, J., McGrew, S., & Ortega, T. (2018). Why Google Can't Save Us: The Challenges of our Post-Gutenberg Moment. In O. Zlatkin-Troitschanskaia, G. Wittum, & A. Dengel (Eds.), *Positive Learning in the Age of Information* (pp. 221–228). Wiesbaden: Springer.
- You, K. H., Lee, S. A., Lee, J. K., & Kang, H. (2013). Why read online news? The structural relationships among motivations, behaviors, and consumption in South Korea. *Information, Communication & Society*, 16(10), 1574–1595. <https://doi.org/10.1080/1369118X.2012.724435>
- Zawacki-Richter, O. (2015). Zur Mediennutzung im Studium – unter besonderer Berücksichtigung heterogener Studierender. *Zeitschrift für Erziehungswissenschaft*, 18(3), 527–549. <https://doi.org/10.1007/s11618-015-0618-6>
- Zlatkin-Troitschanskaia, O., Brückner, S., Molerov, D., & Bisang, W. (2020). *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*. Wiesbaden: Springer.
- Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., Brückner, S. (2019b). Validating a Test for Measuring Knowledge and Understanding of Economics Among University Students. *Zeitschrift für Pädagogische Psychologie*.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019c). On the complementarity of holistic and analytic approaches to performance assessment scoring. *The British Journal of Educational Psychology*. Advance online publication. <https://doi.org/10.1111/bjep.12286>

Appendix

Table 2 Bivariate correlation matrix for media use to inform oneself

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. National Newspapers													
2. Regional Newspapers	.441***												
3. Tabloids	.257***	.419***											
4. Economic Newspapers	.460***	.337***	.340***										
5. Weekly Newspapers	.482***	.358***	.352***	.472***									
6. News Magazines	.334***	.361***	.269***	.313***	.440***								
7. TV News	.228***	.308***	.152***	.223***	.231***	.285***							
8. Economic TV Programs	.264***	.266***	.269***	.314***	.301***	.308***	.395***						
9. Video Platforms	.173**	.085*	.197***	.275***	.191***	.153***	.127***	.227***					
10. Science Journals	.275***	.263***	.290***	.519***	.414***	.275***	.152***	.482***	.254***				
11. Textbooks	.075*	.074*	.100**	.165***	.083*	.125***	.079*	.215***	.158***	.300***			
12. Course Scripts	.004	.059	.056	.113**	.000	.075*	.128***	.092*	.144***	.121**	.456***		
13. Science Data Bases	.153***	.162***	.217***	.175***	.261	.137***	.112*	.314***	.272***	.363***	.249***	.184***	
14. Online Encyclopedias	.152***	.103**	.170***	.224***	.184***	.147***	.141***	.205***	.382***	.234***	.348***	.344***	

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 3 Bivariate correlation matrix for media use to prepare for exams

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. National Newspapers													
2. Regional Newspapers	.855***												
3. Tabloids	.830***	.845***											
4. Economic Newspapers	.826***	.780***	.787***										
5. Weekly Newspapers	.856***	.797***	.824***	.817***									
6. News Magazines	.761***	.747***	.743***	.740***	.834***								
7. TV News	.615***	.615***	.545***	.617***	.628***	.699***							
8. Economic TV Programs	.794***	.780***	.790***	.768***	.785***	.746***	.698***						
9. Video Platforms	.292***	.259***	.257***	.293***	.300***	.274***	.310***	.288***					
10. Science Journals	.781***	.751***	.794***	.808***	.801***	.718***	.546***	.815***	.287***				
11. Textbooks	.219***	.210***	.215***	.290***	.239***	.267***	.214***	.281***	.292***	.309***			
12. Course Scripts	-.321***	-.303***	-.376***	-.249***	-.300***	-.230***	-.106*	-.292***	.151***	-.290***	.338***		
13. Science Data Bases	.634***	.627***	.661***	.605***	.656***	.595***	.438***	.664***	.283***	.700***	.335***	-.186***	
14. Online Encyclopedias	.240***	.245***	.221***	.274***	.270***	.260***	.296***	.264***	.443***	.278***	.373***	.261***	.368***

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 4 Simple and multiple linear regression for media use to inform oneself: Dependent variable = Frequency of national newspaper use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	<i>B</i>	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.053	.019	-.105**	-.053	.02	-.105**
Gender (female = 0)				.372	.108	.132**
Migration background				-.075	.125	-.023**
Parental education				.038	.047	.03
Interest in B&E topics				.398	.075	.2***
Trust in media				.088	.044	.075*
German GPA equivalent				.005	.01	.021
Adjusted R^2		.01**			.076***	.076***
<i>F</i>		7.55			10.48	10.48

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 5 Simple and multiple linear regression for media use to inform oneself: Dependent variable = Frequency of regional newspaper use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.051	.018	-.108**	-.048	.018	-.102**
Gender (female = 0)				.16	.1	.062
Migration background				.231	.114	.076*
Parental education				-.066	.043	-.057
Interest in B&E topics				.199	.07	.109**
Trust in media				.174	.037	.174***
German GPA equivalent				.195	.009	.086*
Adjusted R^2		.01*			.069***	
<i>F</i>		8.09			8.18	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 6 Simple and multiple linear regression for media use to inform oneself: Dependent variable = Frequency of tabloid use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.06	-.016	-.144***	-.033	.015	-.079*
Gender (female = 0)				-.063	.085	-.027
Migration background				.106	.097	.039
Parental education				-.054	.037	-.052
Interest in B&E topics				.105	.059	.064
Trust in media				.327	.033	.35***
German GPA equivalent				.027	.007	.132***
Adjusted R^2		.019***			.152***	
<i>F</i>		14.56			21.59	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 7 Simple and multiple linear regression for media use to inform oneself: Dependent variable = Frequency of economic newspaper use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.054	.017	-.121**	-.049	.017	-.108**
Gender (female = 0)				.487	.092	.196***
Migration background				-.037	.106	-.013
Parental education				.037	.04	.034
Interest in B&E topics				.449	.064	.257***
Trust in media				.081	.038	.076*
German GPA equivalent				.012	.008	.055
Adjusted R^2		.013**			.145***	
<i>F</i>		10.07			17.59	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 8 Simple and multiple linear regression for media use to inform oneself: Dependent variable = Frequency of news magazine use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.07	.02	-.135***	-.072	.02	-.138***
Gender (female = 0)				.292	.111	.101**
Migration background				.141	.127	.042
Parental education				-.039	.047	-.03
Interest in B&E topics				.193	.077	.095**
Trust in media				.245	.042	.218***
German GPA equivalent				.005	.01	.018
Adjusted R^2		.017***			.081***	
<i>F</i>		12.57			9.64	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 9 Simple and multiple linear regression for media use to inform oneself: Dependent variable = Frequency of economic TV magazine use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.042	.016	-.102**	-.033	.016	-.081*
Gender (female = 0)				.235	.088	.103**
Migration background				.0	.1	.0
Parental education				-.055	.038	-.054
Interest in B&E topics				.256	.061	.16***
Trust in media				.091	.034	.101**
German GPA equivalent				.019	.008	.096*
Adjusted R^2		.009**			.07***	
<i>F</i>		7.18			8.34	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 10 Simple and multiple linear regression for media use to inform oneself: Dependent variable = Frequency of video platforms use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.057	.022	-.097*	-.042	.022	-.072
Gender (female = 0)				.556	.12	.172***
Migration background				-.282	.137	-.075*
Parental education				.02	.052	.014
Interest in B&E topics				.226	.084	.099**
Trust in media				.345	.047	.255***
German GPA equivalent				.019	.011	.068
Adjusted R^2		.008*			.135***	
<i>F</i>		6.49			18.95	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 11 Simple and multiple linear regression for media use to inform oneself: Dependent variable = Frequency of science journal use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.046	.137	-.128**	-.04	.0139	-.111**
Gender (female = 0)				.22	.077	.11**
Migration background				-.096	.088	-.041
Parental education				-.007	.033	-.008
Interest in B&E topics				.269	.054	.191***
Trust in media				.075	.03	.095*
German GPA equivalent				.12	.007	.069
Adjusted R^2		.015***			.078***	
<i>F</i>		11.36			9.24	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 12 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of national newspaper use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.048	.024	-.077*	-.044	.025	-.07
Gender (female = 0)				.09	.139	.026
Migration background				-.095	.159	-.023
Parental education				-.069	.06	-.045
Interest in B&E topics				.171	.097	.07
Trust in media				-.008	.057	-.006
German GPA equivalent				-.007	.012	-.023
Adjusted R^2		.004*			.005	
<i>F</i>		4.04			1.45	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 13 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of regional newspaper use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.06	.024	-.096**	-.054	.025	-.086*
Gender (female = 0)				.112	.138	.032
Migration background				-.101	.157	-.025
Parental education				-.13	.06	-.085*
Interest in B&E topics				.123	.096	.05
Trust in media				.075	.051	.056
German GPA equivalent				.0	.012	-.001
Adjusted R^2		.008*			.014*	
<i>F</i>		6.41			2.39	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 14 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of economic newspaper use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.065	.023	-.106**	-.059	.024	-.096*
Gender (female = 0)				.177	.135	.052
Migration background				-.031	.155	-.008
Parental education				-.103	.058	-.068
Interest in B&E topics				.224	.094	.094*
Trust in media				.021	.056	.015
German GPA equivalent				-.002	.012	-.006
Adjusted R^2		.01**			.02*	
<i>F</i>		7.73			2.96	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 15 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of weekly newspaper use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.065	.024	-.106**	-.059	.024	-.096*
Gender (female = 0)				-.084	.136	-.024
Migration background				-.174	.156	-.044
Parental education				-.029	.059	-.019
Interest in B&E topics				.112	.095	.046
Trust in media				.026	.054	.018
German GPA equivalent				-.003	.012	-.01
Adjusted R^2		.01**			.006	
<i>F</i>		7.69			1.56	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 16 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of news magazine use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.062	.024	-.099*	-.059	.025	-.095*
Gender (female = 0)				.0	.138	.0
Migration background				.031	.158	.008
Parental education				-.079	.059	-.051
Interest in B&E topics				.096	.096	.039
Trust in media				.123	.052	.09*
German GPA equivalent				.0	.012	-.001
Adjusted R^2		.008**			.012*	
<i>F</i>		6.73			2.22	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 17 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of TV news use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.076	.025	-.118**	-.079	.025	-.122**
Gender (female = 0)				.017	.14	.005
Migration background				-.058	.161	-.014
Parental education				-.069	.061	-.043
Interest in B&E topics				.159	.098	.063
Trust in media				.262	.057	.175***
German GPA equivalent				.004	.012	.012
Adjusted R^2		.013**			.04***	
<i>F</i>		9.66			5.04	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 18 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of economic TV magazine use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.054	.024	-.087*	-.054	.025	-.087*
Gender (female = 0)				-.06	.135	-.018
Migration background				-.056	.154	-.014
Parental education				-.146	.059	-.096*
Interest in B&E topics				.106	.094	.044
Trust in media				.111	.052	.082*
German GPA equivalent				-.008	.012	-.028
Adjusted R^2		.006*			.017**	
<i>F</i>		5.25			2.7	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 19 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of video platform use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.047	.023	-.077*	-.041	.023	-.066
Gender (female = 0)				.254	.13	.075
Migration background				.0	.149	.0
Parental education				-.05	.056	-.033
Interest in B&E topics				.162	.03	.068
Trust in media				.385	.051	.278***
German GPA equivalent				.007	.011	.022
Adjusted R^2		.005*			.092***	
<i>F</i>		4.06			10.94	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 20 Simple and multiple linear regression for media use to prepare for exam: Dependent variable = Frequency of science journal use ($n = 684$)

Variable	Model 1			Model 2		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
General fluid ability	-.048	.024	-.076*	-.046	.025	-.073
Gender (female = 0)				.03	.138	.009
Migration background				-.138	.158	-.034
Parental education				-.154	.06	-.1*
Interest in B&E topics				.152	.096	.062
Trust in media				.079	.053	.058
German GPA equivalent				-.012	.012	-.04
Adjusted R^2		.004*			.014*	
<i>F</i>		3.96			2.42	

Note. Subsample $n = 709$, missings = 25; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 21 T-tests with BEFKI 11+ scores for media use to inform oneself

	M	SD	95 % CI	$t(df = 343)$	<i>p</i>	<i>d</i>
National Newspapers						
Group 1: Upper Quartile	1.48	1.355	[1.263; 1.687]	2.911	.004**	.314
Group 2: Lower Quartile	1.91	1.398	[1.705; 2.111]			
Regional Newspapers						
Group 1: Upper Quartile	1.19	1.221	[1.003; 1.384]	2.570	.011*	.277
Group 2: Lower Quartile	1.56	1.413	[1.357; 1.767]			
Tabloids						
Group 1: Upper Quartile	.39	.891	[.255; .533]	4.494	>.001***	.485
Group 2: Lower Quartile	.98	1.422	[.772; 1.185]			
Economic Newspapers						
Group 1: Upper Quartile	1.05	1.186	[.865; 1.235]	3.120	.002**	.337
Group 2: Lower Quartile	1.48	1.327	[1.283; 1.668]			

	M	SD	95% CI	<i>t</i> (<i>df</i> = 343)	<i>p</i>	<i>d</i>
Weekly Newspapers						
Group 1: Upper Quartile	1.12	1.215	[.929; 1.308]	2.525	.012*	.273
Group 2: Lower Quartile	1.47	1.315	[1.274; 1.656]			
News Magazines						
Group 1: Upper Quartile	1.44	1.326	[1.321; 1.645]	3.254	.001**	.351
Group 2: Lower Quartile	1.93	1.464	[1.718; 2.142]			
TV News						
Group 1: Upper Quartile	2.48	1.317	[2.269; 2.681]	.366	.714	.040
Group 2: Lower Quartile	2.53	1.437	[2.321; 2.738]			
Economic TV Programs						
Group 1: Upper Quartile	.87	1.077	[.701; 1.037]	2.890	.004**	.312
Group 2: Lower Quartile	1.24	1.298	[1.055; 1.432]			
Video Platforms						
Group 1: Upper Quartile	2.21	1.572	[1.967; 2.458]	2.718	.007**	.293
Group 2: Lower Quartile	2.68	1.619	[2.446; 2.916]			
Science Journals						
Group 1: Upper Quartile	1.7	1.293	[1.498; 1.902]	1.967	.05*	.212
Group 2: Lower Quartile	1.98	1.373	[1.785; 2.183]			
Textbooks						
Group 1: Upper Quartile	.52	.76	[.406; .644]	3.565	>.001***	.385
Group 2: Lower Quartile	.89	1.093	[.733; 1.05]			
Course Scripts						
Group 1: Upper Quartile	3.18	1.207	[2.993; 3.37]	1.608	.109	.174
Group 2: Lower Quartile	3.4	1.247	[3.214; 3.575]			

	M	SD	95 % CI	<i>t</i> (df = 343)	<i>p</i>	<i>d</i>
Science Databases						
Group 1: Upper Quartile	.69	1.167	[.505; .87]	2.527	.012*	.273
Group 2: Lower Quartile	1.02	1.272	[.837; 1.206]			
Online Encyclopedias						
Group 1: Upper Quartile	2.23	1.309	[2.027; 2.436]	1.535	.126	.166
Group 2: Lower Quartile	2.46	1.433	[2.252; 2.667]			

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 22 T-tests with BEFKI scores for media use to prepare for exam

	M	SD	95 % CI	<i>t</i> (df = 343)	<i>p</i>	<i>d</i>
National Newspapers						
Group 1: Upper Quartile	1.24	1.779	[.96; 1.515]	1.453	.147	.157
Group 2: Lower Quartile	1.51	1.678	[1.265; 1.752]			
Regional Newspapers						
Group 1: Upper Quartile	1.09	1.758	[.819; 1.368]	2.337	.02*	.252
Group 2: Lower Quartile	1.53	1.741	[1.283; 1.788]			
Tabloids						
Group 1: Upper Quartile	.94	1.785	[.665; 1.222]	1.404	.161	.152
Group 2: Lower Quartile	1.21	1.743	[.958; 1.464]			
Economic Newspapers						
Group 1: Upper Quartile	1.17	1.716	[.901; 1.437]	2.234	.026*	.241
Group 2: Lower Quartile	1.58	1.724	[1.334; 1.834]			
Weekly Newspapers						
Group 1: Upper Quartile	1.15	1.763	[.875; 1.425]	2.467	.014*	.266
Group 2: Lower Quartile	1.61	1.7	[1.364; 1.857]			

	M	SD	95% CI	<i>t</i> (<i>df</i> = 343)	<i>p</i>	<i>d</i>
News Magazines						
Group 1: Upper Quartile	1.31	1.816	[1.029; 1.596]	2.113	.035*	.228
Group 2: Lower Quartile	1.71	1.706	[1.466; 1.961]			
TV News						
Group 1: Upper Quartile	1.62	1.836	[1.332; 1.905]	2.593	.01**	.28
Group 2: Lower Quartile	2.13	1.816	[1.866; 2.393]			
Economic TV Programs						
Group 1: Upper Quartile	1.17	1.792	[.889; 1.449]	1.759	.08	.19
Group 2: Lower Quartile	1.5	1.729	[1.252; 1.753]			
Video Platforms						
Group 1: Upper Quartile	2.35	1.778	[2.072; 2.627]	1.595	.112	.172
Group 2: Lower Quartile	2.64	1.636	[2.406; 2.881]			
Science Journals						
Group 1: Upper Quartile	1.15	1.816	[.866; 1.434]	1.109	.268	.12
Group 2: Lower Quartile	1.36	1.733	[1.111; 1.614]			
Textbooks						
Group 1: Upper Quartile	2.43	1.814	[2.148; 2.715]	1.023	.307	.11
Group 2: Lower Quartile	2.62	1.641	[2.384; 2.860]			
Course Scripts						
Group 1: Upper Quartile	3.38	1.678	[3.119; 3.643]	.044	.965	.005
Group 2: Lower Quartile	3.39	1.648	[3.15; 3.628]			
Science Data Bases						
Group 1: Upper Quartile	1.31	1.867	[1.021; 1.604]	1.341	.181	.145
Group 2: Lower Quartile	1.58	1.81	[1.316; 1.841]			

	M	SD	95% CI	$t(df = 343)$	p	d
Online Encyclopedias						
Group 1: Upper Quartile	2.5	1.602	[2.25; 2.750]	.874	.383	.094
Group 2: Lower Quartile	2.65	1.658	[2.413; 2.895]			

* $p < .05$; ** $p < .01$; *** $p < .001$



3.3

Multiple Document Comprehension of University Students

Test Development and Relations to Person and Process Characteristics

Schoor, C., Hahnel, C., Mahlow, N., Klagges, J., Kroehne, U., Goldhammer, F., and Artelt, C.

Abstract

Multiple document comprehension is the ability to construct an integrated representation of a specific topic based on several sources. It is an important competence for university students; however, there has been so far no established instrument to assess multiple document comprehension in a standardized way. Therefore, we developed a test covering four theory-based cognitive requirements: The corroboration of information across texts, the integration of information across texts, the comparison of sources and source evaluations across texts, and the comparison of source-content links across texts. The developed test comprised 174 items and was empirically examined in a study with 310 university students. Several items had to be excluded due to psychometric misfit and differential item functioning. The resulting final test contains 67 items within 5 units (i.e., test structures of 2–3 texts and related items) and has been shown to fit a unidimensional IRT Rasch model. The test score showed expected relationships to the final school exam grade, the study level (Bachelor/Master), essay performance, sourcing behavior, as well as mental load and mental effort.

Keywords

Multiple document comprehension, university students, assessment, multiple document literacy, sourcing, mental load, mental effort, basic computer skills, reading frequency

1 Introduction

To fulfill the requirements of their studies and their later job positions, university students have to be able to familiarize themselves with specific topics. Often, they are confronted with several sources and text documents from which they have to select and integrate relevant information across documents while taking characteristics of the source into account. As such, these requirements exceed the requirements of comprehending single texts.

It is expected of German students entering the university that they have basic skills of multiple document comprehension (Kultusministerkonferenz 2012). Students leaving the university with a master's degree are supposed to be able to integrate knowledge, to deal with complexity, and to independently acquire new knowledge (Arbeitskreis Deutscher Qualifikationsrahmen 2011). We refer to the underlying ability of students as *multiple document comprehension* (MDC) and define it as the ability to construct an integrated representation of a specific topic based on several different sources of information (Anmarkrud et al. 2014; Schoor et al. 2020).

However, many university students do not meet these expectations (Gruenbaum 2012; Peter 2019; Rouet et al. 1997). Nevertheless, there is no standardized test so far that covers several theory-based aspects of MDC (Section 2) as a competence and assesses it in a valid way. To close this gap, we developed a computer-based test of MDC for university students. This test is based on the assumption that in the context of their university studies, students encounter mainly trustworthy documents on topics for which they often possess little prior knowledge. This is a situation typical for learning with text at the university, when students receive documents from their university lecturer or find them in a scientific database, for example. Therefore, the test does not cover competences that relate to the search and selection of documents or to the dealing with information that contradicts prior notions. In the present chapter, we introduce the framework of the developed test and present results on the relationship of MDC as a competence with person characteristics and characteristics of the process of dealing with multiple documents.

2 Multiple Document Comprehension of University Students

Understanding multiple documents of a specific topic necessitates not only the understanding of single texts (e.g., Kintsch 1998), but also, for example,

1. to notice that there are different perspectives on the same topic and to relate these perspectives to each other (Britt et al. 1999; Britt and Rouet 2012),
2. to integrate information across documents (Britt and Sommer 2004; Cerdán and Vidal-Abarca 2008; Gil et al. 2010b),
3. to recognize and deal with conflicting information (Bråten et al. 2014; Keck et al. 2015; Maier and Richter 2013; Stadtler and Bromme 2014),
4. to notice biases of sources and take them into account to make own decisions (Braasch et al. 2012; Bråten et al. 2016; Britt and Aglinskas 2002; Kammerer and Gerjets 2014; Kammerer et al. 2016),
5. to compare and corroborate information from different sources (Rouet et al. 1997; Wineburg 1991).

Although these requirements often occur in the context of multiple documents, it is probably not the existence of multiple documents but multiple sources that makes the difference. By the term *document* we refer to the structural entity transmitting content; that is a journal article, a web page, or a textbook chapter, for example. With the term *source* we refer to the originator of the content. That means that in different journal articles (documents) different authors (sources) can present different models about a topic (content). The same models could also be explained in a textbook chapter (document), thus presenting content from multiple sources. Of course, also the textbook chapter has a source (its author). In contrast to multiple documents, however, a well-written textbook chapter is supposed to support the reader in the task of integrating and comparing the content of different sources (Goldman and Scardamalia 2013). That is, we expect the authors of textbook chapters to describe commonalities and differences of the different models that they are presenting.

2.1 Theoretical Approaches to Multiple Document Comprehension

There are several theoretical frameworks that can be applied to multiple document comprehension. For example, multiple document comprehension could be considered a special case of understanding multiple representations (Ainsworth 2006; Schnotz and Bannert 2003; Seufert 2009), a special task embedded in a context (RESOLV: Rouet et al. 2017), or guided by interest, attitudes, and the availability of MDC strategies (CAEM: List and Alexander 2017). For the development of a standardized test, we refer to the Documents Model Framework (e.g., Britt and Rouet 2012) and the strategies identified by Wineburg (1991). Both approaches specify representations and strategies from which cognitive requirements can be derived that represent MDC.

The Documents Model Framework (e.g., Britt and Rouet 2012) specifies the adequate cognitive representation of multiple documents. In this framework, it is supposed that beyond single text representation a so-called *documents model* has to be constructed. The documents model is a mental model consisting of two sub-models: the *integrated situation model* and the *intertext model*. The integrated situation model is an integrated representation of the content of the multiple documents and prior knowledge. The intertext model is a representation of meta-information about the sources like author, form, goals, or cultural background of the documents (Perfetti et al. 1999). The different sources are related to each other with predicates like “contradicts” or “confirms”. In addition, the intertext model and the integrated situation model are connected with each other such that the most central information in the integrated situation model is tagged with source information (e.g., Britt and Rouet 2012).

With regard to strategies that are beneficial for multiple document comprehension, Wineburg (1991) found that experts – in contrast to novices – engaged more in behaviors of corroboration, contextualization, and sourcing. That is, experts compared information across documents, they related the information from the documents to their prior knowledge, and they took information about the source into account very early in their processing of a document. While contextualization is already relevant for the construction of a situation model of a single text, corroboration helps developing an integrated situation model, while sourcing is needed to construct an intertext model.

2.2 Prior Approaches to Assessing Multiple Document Comprehension

In prior research, multiple document comprehension has often been assessed by means of essays or intertext inference verification tasks (Primor and Katzir 2018). Essays are considered an expressive task, in which the participants have to write a text about multiple documents. The essays are usually manually coded holistically or with regard to a specific aspect (e.g., sourcing, integration) (e.g., Britt and Aglinskas 2002; Kammerer and Gerjets 2014; Rouet et al. 1997; Stadler et al. 2014). Therefore, essays are a time-consuming way of assessing MDC. Moreover, it seems that they might be measuring not only multiple document comprehension but also writing skills (Griffin et al 2017; Primor and Katzir 2018).

A receptive approach to assessing multiple document comprehension is the use of intertext inference verification tasks (e.g., Braasch et al. 2014; Salmerón et al. 2010; see also Maier and Richter 2013; Schmalhofer and Glavanov 1986). In these tasks, the integrated situation model is targeted. The participants have to judge whether a given statement is a valid inference given the texts read. Although this approach is economic and objective, only one specific aspect of multiple document comprehension – the integrated situation model – is assessed, but not the intertext model or the whole documents model.

2.3 Characteristics of Multiple Document Comprehension in the University Context

The present test of multiple document comprehension was developed for the university context. During the search for information in the internet, the trustworthiness of documents plays a major role and has therefore been focused by recent research (e.g., Braasch and Bråten 2017; Bråten et al. 2016; Britt and Aglinskas 2002; Kammerer and Gerjets 2014; Paul et al. 2017; Scharrer and Salmerón 2016; Strømsø et al. 2013). However, at the university, students usually have to read multiple documents that have been given to them by their university lecturer or that they found in a scientific database. Therefore, the trustworthiness of documents usually is not a major issue (which is different in learning from information found on the Internet), but only whether the documents can contribute to the question one tries to answer. The search and selection of documents can be considered a competence independent from MDC (information literacy, see DBV 2009; Homann 2000; Lau 2006). Therefore, the test that we developed focuses on the *comprehension* of multiple documents, leaving out the search and selection of documents.

Although there are differences across disciplines with regard to the characteristics of multiple documents, there are commonalities that justify to assume MDC to be a generic cross-disciplinary competence (Goldman et al. 2016). Consequently, the present test has been developed to cover the generic MDC competence. There is also evidence that MDC can develop during the course of students' university studies (Britt and Aglinskias 2002; Mühlen et al. 2016), which is aimed to be reflected in the test score.

3 Development of the Multiple Document Comprehension Test

The development of our test was based on the documents model (Britt and Rouet 2012) and Wineburg's (1991) strategies. Moreover, the items of the test should not be solvable correctly with only one text. For the sake of simplicity, we assumed each document to have and report on only one source, so that multiple sources implied multiple documents.

3.1 Framework: Cognitive Requirements

The items of the test were based on four cognitive requirements that were derived from the documents model (Britt and Rouet 2012) and Wineburg's (1991) strategies:

1. Corroboration of information across texts. Items with this requirement necessitate that the reader compares information stemming from different texts. Either this information can be found directly in the text or a simple inference is necessary. This cognitive requirement is derived from Wineburg's (1991) strategy of corroboration.
2. Integration of information across texts. This requirement refers to the integrated situation model. For solving items with this requirement correctly, information from different documents has to be combined. For easier items, this combination might be additive; for more difficult items, it is necessary to draw an inference based on information stemming from different texts.
3. Comparison of sources and source evaluations across texts. This requirement necessitates an intertext model. Items require the judgment of single sources (i.e. sourcing) and their comparison across texts/sources.

4. Comparison of source-content links across texts. Items with this requirement necessitate a documents model. Content has to be represented together with its source (source-content links) and compared across texts.

3.2 The Developed Test

We developed six units on different topics to cover different disciplines (Table 1). A unit consisted of two or three texts on the topic and up to 15 items, each of which covered one of the four abovementioned cognitive requirements. In two units, an essay task had to be completed before access was granted to the other items. Due to their expressive nature, the essays were not included in the test but used as a criterion for validation. The content of the units was fictitious, except for the unit “Universe”, to reduce effects of prior knowledge. For each text, information about the source was also provided. The items had to be answered either in a dichotomous verification format (e.g., true/false) or in a single choice format (1 out of 4).

Each unit started with an overview of the number of texts and items. On this page, a reading goal was set as well (Gil et al. 2010a; Stadler et al. 2014). If there was an essay task in the unit, the task referred to this reading goal. The computer-based test was implemented with the CBA ItemBuilder (Rölke 2012). Figure 2 shows a screenshot of the test with its functionalities. All functionalities were explained to the participants in a video-based tutorial before the actual test was taken.

3.3 Empirical Examination of the Test

To empirically examine the MDC test (Schoor et al. 2020), a convenience sample of 310 university students of the humanities and social sciences from two German universities of either the first Bachelor’s or Master’s semester worked on three out of the six units of the test. Using a balanced incomplete block design and an additional rotation of the units, 60 different testlets were created to counterbalance each possible pair of units and the order of units. The students were randomly assigned to one of these testlets. In addition to the MDC test, the participants worked on several other scales such as reading frequency, basic computer skills, mental load and mental effort, epistemic beliefs, and goal orientations. The three units of the MDC test took about 1 to 1.5 hours; the whole session about 2 hours.

The scored items were tested with regard to their fit to the Rasch model. Due to misfit and differential item functioning, several items and the whole unit “forgiv-

ing” had to be excluded (Table 1). The final test encompasses 67 items in five units. As a person parameter for MDC, we used weighted likelihood estimates (WLE scores; Warm 1989), which had an acceptable WLE reliability of .67.

In addition, we tested whether the resulting test was unidimensional. The unidimensional model was tested against a four-dimensional model representing the different cognitive requirements of the items, against a five-dimensional model representing the unit structure, and against a Rasch testlet model with a general factor and five subdimensions representing the unit structure. The results favored the unidimensional model (for more details, see Schoor et al. 2020).

Table 1 Overview over the units of the MDC test by Schoor et al. (2020)

Unit	Content	Multiple choice format				Verification format	
		Number of items of type...				Number of items of type...	
		1	2	3	4	1	3
Nothing	2 reviews of the fictitious novel “Nothing” which were allegedly published in newspapers, both describing content and quality of the novel, inspired by the novel “Nada” by Carmen Laforet	1 (1)	8 (2)	1 (0)	2 (2)	16 (4)	8 (4)
Universe	3 popular science texts on the end of the universe from a physical-cosmological perspective; the texts encompass scenarios of the end, a description of forces, and a report of new empirical data	-	5 (4)	1 (1)	3 (3)	8 (7)	-
Catalano	2 short biographies on the life of the (fictitious) mafia boss Catalano, inspired by the life of Al Capone	-	11 (6)	2 (2)	1 (1)	8 (2)	-

Unit	Content	Multiple choice format				Verification format	
		Number of items of type...				Number of items of type...	
		1	2	3	4	1	3
Forgiving	3 texts imitating textbook chapters, each presenting a (fictitious) theoretical model on forgiving; the models differ with regard to their general approach (process model, influencing factors, dimensions)	-	5 (0)	-	6 (0)	8 (0)	16 (0)
2134	3 texts on a “historical” event in the year 2134: the arrival of extraterrestrials on earth. The texts are written as if they were historical documents: One text is an internal report of an observatory, one an internal governmental report, and one is a political speech	-	8 (4)	1 (0)	3 (3)	8 (3)	8 (1)
Animals	3 textbook texts each presenting one (fictitious) literature studies approach on how to interpret animals in novels; like the texts in “Forgiving”, the texts present structurally different approaches	-	4 (2)	-	8 (5)	8 (2)	16 (8)

Note. Type of items refers to cognitive requirements of the items (1 = corroboration of information across texts; 2 = integration of information across texts; 3 = comparison of sources and source evaluation across texts; 4 = comparison of source-content links across texts). Verification format was used only for items of type 1 or 3. In brackets: Number of items left in the final test.

access to documents access to items access to source information possibility to comment on the margin time spent so far exit

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 Bisheriger Zeitaufwand 11 Minuten

Meine Lösung zu Aufgabe 1

Stellen Sie sich vor, Sie müssten in einem Seminar an der Universität ein Referat über Verzeihen halten und hätten nur die drei Texte zur Verfügung. Schreiben Sie eine Gliederung Ihres Referats (mit Nennung von Inhaltspunkten) und begründen Sie die Bedeutung für die 15 Punkte.

1. Verzeihen als psychologisches Phänomen (Einführung, Definition)

2. Verzeihen beschreiben

2.1 Kognitive, emotionale und Verhaltensaspekte beim Verzeihen

2.2 Prozessmodelle nach Shavelson & van der Beekle (2012)

2.3 Operationalisierung von Verzeihen (Dürl & Henrich, 2006; Sulth et al., 2006)

3. Einflüsse auf das Verzeihen

3.1 Personelle Einflüsse (Geschlecht, Kontakt zum Täter)

3.2 Situationale Einflüsse (z.B. Schwere der Tat, kulturelle & religiöse Einflüsse?)

3.4 Differenzierung Selbst- und Fremdverzeihen

4. Zusammenfassung

Aufgabe im Text

Schwerpunkt der Studie

mandantiertes Interview?

access to documents access to items access to source information possibility to comment on the margin time spent so far exit

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 Bisheriger Zeitaufwand 11 Minuten

Meine Lösung zu Aufgabe 1

Stellen Sie sich vor, Sie müssten in einem Seminar an der Universität ein Referat über Verzeihen halten und hätten nur die drei Texte zur Verfügung. Schreiben Sie eine Gliederung Ihres Referats (mit Nennung von Inhaltspunkten) und begründen Sie die Bedeutung für die 15 Punkte.

1. Verzeihen als psychologisches Phänomen (Einführung, Definition)

2. Verzeihen beschreiben

2.1 Kognitive, emotionale und Verhaltensaspekte beim Verzeihen

2.2 Prozessmodelle nach Shavelson & van der Beekle (2012)

2.3 Operationalisierung von Verzeihen (Dürl & Henrich, 2006; Sulth et al., 2006)

3. Einflüsse auf das Verzeihen

3.1 Personelle Einflüsse (Geschlecht, Kontakt zum Täter)

3.2 Situationale Einflüsse (z.B. Schwere der Tat, kulturelle & religiöse Einflüsse?)

3.4 Differenzierung Selbst- und Fremdverzeihen

4. Zusammenfassung

Aufgabe im Text

Schwerpunkt der Studie

mandantiertes Interview?

Kapitel 10: Verzeihen

Bertrik Mahanna Gandhi bezeichnete Verzeihen als eine Eigenschaft des Starken, weil dem Schwachen die Kraft und der Mut zum Verzeihen fehle. Heute besteht ein allgemeiner Konsens darüber, dass Verzeihen ein Prozess ist, der sich in einer prozessualen Veränderung von Gefühlen, Gedanken und Verhalten gegenüber einem Übeltäter äußert. Verzeihen geschieht bewusst und bedingungslos. Gleichzeitig ist Verzeihen durch personale wie kulturelle Gegebenheiten beeinflusst. Dabei wird unterschieden zwischen „anderen verzeihen“ und „sich selbst verzeihen“.

10.1 Interviewstudie von Thomsen et al. (1985)

Wenn Verzeihen von vielen verschiedenen inneren und äußeren Faktoren abhängig ist, welche Faktoren sind dies? Mit dieser Frage beschäftigte sich 1985 ein amerikanisches Forscherteam. Im Bestreben, die Einflussfaktoren von Selbst- und Fremdverzeihen zu ermitteln und zu untersuchen, wurden an der Georgia State University in Atlanta Interviews von befreundeten Studierenden durchgeführt. Die Aufgabe der Studierenden bestand darin, einen in der Vergangenheit liegenden Konflikt zwischen den beiden Interviewten zu beschreiben und über ihre Gedanken

und Strategien des Umgangs mit dem Konflikt zu sprechen. Das Hauptziel der Untersuchung lag darin herauszufinden, welche Faktoren Verzeihen oder Selbstverzeihen beeinflussen konnten. Insgesamt wurden dreißig Zweiergruppen interviewt. Die Interviewer orientierten sich dabei an einem vorab entwickelten Leitfaden. Während neunzehn Paare einen Konflikt mit anschließender Problemlösung beschrieben, zeichnete sich in elf der Interviews ein nur teilweise bis gar nicht gelöster Konflikt ab.

Im Laufe der Interviews konnten verschiedene Faktoren ermittelt werden, die Verzeihen und Selbstverzeihen beeinflussen konnten. Thomsen et al. (1985) haben diese jeweils an die folgenden drei Dimensionen zusammengefasst:

Figure 2 Screenshot of the computer-based test with functionalities, especially access to source information, highlighting, and commenting

4 Relations of Multiple Document Comprehension to Person and Process Characteristics

Multiple document comprehension is supposed to be meaningfully related to person characteristics and characteristics of the comprehension process. To examine the validity of the MDC test score interpretation, we related the MDC test score to these variables. By person characteristics, we refer to demographic variables, basic computer skills, and reading frequency. With regard to process characteristics, we look at the way in which the participants build their understanding of multiple documents, as indicated by sourcing and mental effort, for example. We developed specific expectations about how these variables should be related to the MDC test score (see below). All hypotheses were tested with the same sample as described above.

4.1 Multiple Document Comprehension and Demographic and Study-Related Variables

Students entering the university are supposed to possess basic skills in MDC. Since this is an ability that should have been taught in school, the final school exam (German Abitur) grade should be related to the MDC test score such that the better the final school exam grade, the better the MDC test score. Moreover, since MDC should develop during their university studies, we expected Master's students to have a higher MDC test score than Bachelor's students, and we expected the final school exam grade to be more closely related to the MDC score of Bachelor's than of Master's students.

The final school exam grade correlated in the expected direction with the MDC test score ($r = -.44, p < .001$) and Master's students achieved higher MDC scores than Bachelor's students ($\beta_{\text{standardized}} = 0.22, p < .001$) (Schoor et al. 2020). The correlation of the MDC test score with the final school exam grade was descriptively higher for Bachelor's ($r = -.44, p < .001$) than for Master's students ($r = -.35, p < .001$), but this difference was not statistically significant (Schoor et al. 2020). A t test on the MDC test scores revealed no significant gender differences ($M_{\text{male}} = 0.09, SD_{\text{male}} = 0.74; M_{\text{female}} = -0.01, SD_{\text{female}} = 0.75; t_{308} = 0.98, p = .33$).

4.2 Multiple Document Comprehension and Essay Performance

Prior research has often operationalized MDC by means of essay performance. However, essay performance also includes expressive abilities. Therefore, performance in an essay task was expected to correlate only in a moderate way with the MDC test score. In our study, an essay task was included in the units “Universe” and “Nothing”. Each essay was scored with regard to the mentioning of ideas central to answering the task (for scoring procedure and scheme, see Schoor et al. 2020). The MDC test score correlated with the essay score in the unit “Universe” at $r = .32$ ($p < .001$) and with the essay score in the unit “Nothing” at $r = .44$ ($p < .001$) (Schoor et al. 2020).

4.3 Multiple Document Comprehension and Sourcing Behavior

As a central MDC-related strategy (Wineburg 1991), sourcing is supposed to be closely related to multiple document comprehension. Since our MDC test is computer-based, log data are available that allow analyzing the participants’ behavior. To validate different indicators of sourcing behavior, Hahnel, Kroehne et al. (2019) investigated the following indicators of sourcing:

1. Proactive sourcing. Referring to sourcing as a heuristic that provides a framework for the processing of documents, sourcing was operationalized as a dichotomous indicator of whether or not the source information of a document was accessed within the first 10 % of overall document reading time.
2. Repeated sourcing. Sourcing might be necessary to update source information in the working memory. This might be triggered, for example, by conflicting information (Braasch et al. 2012). Accordingly, this indicator was operationalized as dichotomous indicator of whether or not the source information of a document was accessed more than once.
3. Task-related sourcing. In the MDC test, there were items that necessitated source evaluation. These items might have triggered sourcing. Therefore, task-related sourcing was a dichotomous indicator of whether or not after accessing an item the source information of a document was accessed with a maximum of 10 seconds time on the corresponding document.
4. A general indicator of whether or not the source information of a document was assessed.

All these indicators were positively and significantly related to the MDC test score ($b = 0.27 - 0.53$, all $ps < .001$ except for proactive sourcing: $p < .05$). Moreover, they were also expected to relate to unit characteristics and properties of the test administration in a specific way. With regard to the validity of their interpretation, the indicators of proactive sourcing and repeated sourcing showed the expected patterns of relationships with other variables such as MDC test score and graduation grade. Therefore, it can be concluded that these indicators reflect their intended meaning well (for details, see Hahnel, Kroehne et al. 2019). This was not the case for task-related sourcing and therefore the use of this indicator was not recommendable. The general indicator revealed to be a mixture of the other three sourcing indicators (Hahnel, Kroehne et al. 2019).

4.4 Multiple Document Comprehension and Mental Load and Mental Effort

Multiple document comprehension demands a lot from readers. It can be assumed that to develop a documents model, the reader has to hold and process multiple interacting elements in working memory. This can result in high cognitive load (Sweller 2010). Mental load (the perceived task difficulty) and mental effort can be considered two components of cognitive load (Paas 1992), which we assessed together with the MDC test. The participants filled in a questionnaire on mental load and mental effort (Krell 2015) after each MDC test unit. Amongst others, we expected that mental load would increase with the number of documents in the unit and their total length, which we did not expect to be the case for mental effort since this measure should be dependent on the students' individual engagement. With regard to MDC, we expected the probability to solve the items correctly to be negatively related to mental load and positively related to mental effort.

Instead of the MDC test score, the scores of each item were predicted using generalized linear mixed models (Hahnel, Schoor et al. 2019). As hypothesized, mental load increased with the number of documents ($b = 0.99, p < .001$). Mental effort was not significantly related to the number of documents ($b = -0.01, p > .05$). As also expected, mental load was negatively ($b = -0.13, p < .001$) and mental effort positively ($b = 0.18, p < .001$) related to the probability to solve the MDC tasks correctly (for more detailed results, see Hahnel, Schoor et al. 2019).

4.5 Multiple Document Comprehension and Basic Computer Skills

Basic computer skills are defined as „the fundamental ability and speed of performing basic actions in graphical user interfaces of computers to access, collect, and provide information“ (Goldhammer et al. 2013). Since our MDC test was computer-based, we also assessed basic computer skills to examine whether the MDC test score cannot be explained merely by basic computer skills. We used a further developed and short version of the test by Goldhammer et al. (2013), encompassing five tasks (i.e., forward an e-mail, online banking transaction request, send an e-mail out of a text processing software, literature search in a data base, search for a flight). Each task was scored dichotomously (solved correctly vs. not solved correctly), and a sum score was built. This sum score of basic computer skills was positively correlated with the MDC test score ($r = .25, p < .001$). However, the internal consistency of the scale was not acceptable (McDonald's $w = .28$) and too low to meaningfully interpret the relationship with the MDC test score.

4.6 Multiple Document Comprehension and Reading Frequency

Reading frequency has been shown to be related to (single text) reading comprehension (e.g., Guthrie et al. 1999; Locher and Pfof 2019). Since it can be assumed that single text reading comprehension is a necessary but not a sufficient prerequisite for multiple document comprehension, we expected reading frequency to be positively related to the MDC test score. Reading frequency was measured by means of five items from the NEPS study (Blossfeld et al. 2011) asking for the daily amount of reading for study purposes and for leisure as well as how often the participants read fictional texts, non-fiction texts, instructions, advertisements, or commenting texts in their leisure time. Table 2 displays the findings. Except for the amount of reading commenting texts during leisure time, no significant relationship of reading frequency with MDC was found.

Table 2 Correlations of reading frequency items with MDC

Variables of Reading Frequency		<i>r</i>
Daily amount	Reading for studies	-.07
	Reading in leisure time	.03
Genre-specific reading	fictional texts	.06
	non-fiction texts	-.01
	instructions	-.03
	advertisements	-.01
	commenting texts	.13*

Note. * $p < .05$.

5 Conclusion and Outlook

Our aim was to develop a test of multiple document comprehension that assesses the generic, domain-independent ability of university students to deal with multiple documents on a topic for which they have low or no prior knowledge. The results of the first empirical evaluation of the test showed that the developed test is objective, reliable, and valid. The test format in terms of closed-ended items makes the assessment objective and economic. The reliability is lower than optimal, but still acceptable. The relationships between the MDC test score and other variables indicate that the test measures what it is supposed to measure. Thus, the interpretation of the test score as reflecting the ability of university students to create a comprehensive mental representation from multiple documents by corroborating and integrating information across texts, comparing and evaluating their respective sources, and creating and comparing source-content links across documents can be considered valid at present.

Currently, further studies are in preparation. For one, we conducted a second study in which we included measures of (single text) reading competence and working memory to analyze their relationship to MDC. Further contributions will cover the relationship of MDC with the task model (Rouet et al. 2017), goal orientations (Spinath et al. 2002), and epistemic beliefs (e.g., Ferguson 2015). Based on the gathered log data, strategies that the participants apply also will be examined in future research. Further research is required, for example, to investigate the longitudinal development of MDC over the course of university studies or to expand the generalizability of the present results to students of all subjects.

Funding

The project on which this report is based was funded by the German Federal Ministry of Education and Research, funding code 01PK15008. Responsibility for the contents of this publication rests with the authors.

References

- Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), pp. 183–198. doi: 10.1016/j.learninstruc.2006.03.001
- Anmarkrud, Ø., Bråten, I., & Strømsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences, 30*, pp. 64–76. doi: 10.1016/j.lindif.2013.01.007
- Arbeitskreis Deutscher Qualifikationsrahmen (2011). *Deutscher Qualifikationsrahmen für lebenslanges Lernen*. Retrieved from http://www.dqr.de/media/content/Der_Deutsche_Qualifikationsrahmen_fue_lebenslanges_Lernen.pdf
- Blossfeld, H.-P., Roßbach, H.-G., & Maurice, J. v. (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaften, 14*.
- Braasch, J. L. G., & Bråten, I. (2017). The discrepancy-induced source comprehension (D-ISC) model: Basic assumptions and preliminary evidence. *Educational Psychologist, 52*(3), pp. 167–181. doi: 10.1080/00461520.2017.1323219
- Braasch, J. L. G., Bråten, I., Strømsø, H. I., & Anmarkrud, Ø. (2014). Incremental theories of intelligence predict multiple document comprehension. *Learning and Individual Differences, 31*, pp. 11–20. doi: 10.1016/j.lindif.2013.12.012
- Braasch, J. L. G., Rouet, J.-F., Vibert, N., & Britt, M. A. (2012). Readers' use of source information in text comprehension. *Memory & Cognition, 40*(3), pp. 450–465. doi: 10.3758/s13421-011-0160-6
- Bråten, I., Salmerón, L., & Strømsø, H. I. (2016). Who said that? Investigating the Plausibility-Induced Source Focusing assumption with Norwegian undergraduate readers. *Contemporary Educational Psychology, 46*, pp. 253–262. doi: 10.1016/j.cedpsych.2016.07.004
- Bråten, I., Ferguson, L. E., Strømsø, H. I., & Anmarkrud, Ø. (2014). Students working with multiple conflicting documents on a scientific issue: Relations between epistemic cognition while reading and sourcing and argumentation in essays. *British Journal of Educational Psychology, 84*(1), pp. 58–85. doi: 10.1111/bjep.12005
- Britt, M. A., & Rouet, J.-F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276–314). New York: Cambridge University Press.
- Britt, M. A., & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Reading Psychology, 25*(4), pp. 313–339. doi: 10.1080/02702710490522658

- Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*(4), pp. 485–522. doi: 10.1207/s1532690x-ci2004_2
- Britt, M. A., Perfetti, C. A., Sandak, R., & Rouet, J.-F. (1999). Content integration and source separation in learning from multiple texts. In S. R. Goldman, A. C. Graesser, & P. Van den Broek (Eds.), *Narrative, comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 209–233). Mahwah, NJ: Erlbaum.
- Cerdán, R., & Vidal-Abarca, E. (2008). The effects of tasks on integrating information from multiple documents. *Journal of Educational Psychology, 100*(1), pp. 209–222. doi: 10.1037/0022-0663.100.1.209
- DBV (2009). *Standards der Informationskompetenz für Studierende*. Retrieved from www.bibliotheksverband.de/fileadmin/user_upload/Kommissionen/Kom_Dienstleistung/Publikationen/Standards_Infokompetenz_03.07.2009_endg.pdf
- Ferguson, L. E. (2015). Epistemic beliefs and their relation to multiple-text comprehension: A Norwegian program of research. *Scandinavian Journal of Educational Research, 59*(6), pp. 731–752. doi: 10.1080/00313831.2014.971863
- Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. I. (2010a). Summary versus argument tasks when working with multiple documents: Which is better for whom? *Contemporary Educational Psychology, 35*(3), pp. 157–173. doi: 10.1016/j.cedpsych.2009.11.002
- Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. I. (2010b). Understanding and integrating multiple science texts: Summary tasks are sometimes better than argument tasks. *Reading Psychology, 31*(1), pp. 30–68. doi: 10.1080/02702710902733600
- Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills. Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment, 29*(4), pp. 263–275. doi: 10.1027/1015-5759/a000153
- Goldman, S. R., & Scardamalia, M. (2013). Managing, understanding, applying, and creating knowledge in the information age: Next-generation challenges and opportunities. *Cognition and Instruction, 31*(2), pp. 255–269. doi: 10.1080/10824669.2013.773217
- Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., . . . Project, R. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist, 51*(2), pp. 219–246. doi: 10.1080/00461520.2016.1168741
- Griffin, T. D., Wiley, J., Britt, M. A., & Salas, C. R. (2017). The role of clear thinking in learning science from multiple-document inquiry tasks. *International Electronic Journal of Elementary Education, 5*(1), pp. 63–78.
- Gruenbaum, E. A. (2012). Common literacy struggles with college students: Using the reciprocal teaching technique. *Journal of College Reading and Learning, 42*(2), pp. 109–116. doi: 10.1080/10790195.2012.10850357
- Guthrie, J. T., Wigfield, A., Metsala, J. L., & Cox, K. E. (1999). Motivational and cognitive predictors of text comprehension and reading amount. *Scientific Studies of Reading, 3*(3), pp. 231–256.
- Hahnel, C., Kroehne, U., Goldhammer, F., Schoor, C., Mahlow, N., & Artelt, C. (2019). Validating process variables of sourcing in an assessment of multiple document comprehension. *British Journal of Educational Psychology, 89*(3), pp. 524–537. doi: 10.1111/bjep.12278

- Hahnel, C., Schoor, C., Kröhne, U., Goldhammer, F., Mahlow, N., & Artelt, C. (2019). The role of cognitive load for university students' comprehension of multiple documents. *Zeitschrift für Pädagogische Psychologie*, 33(2), pp. 105–118. doi: 10.1024/1010-0652/a000238
- Homann, B. (2000). Das Dynamische Modell der Informationskompetenz (DYMIK) als Grundlage für bibliothekarische Schulungen. In G. Knorz & R. Kuhlen (Eds.), *Informationskompetenz – Basiskompetenz in der Informationsgesellschaft. Proceedings des 7. Internationale Symposiums für Informationswissenschaft (ISI 2000), Darmstadt, 8.–10. November 2000* (pp. 195–206). Konstanz: UVK Verlag.
- Kammerer, Y., & Gerjets, P. (2014). Quellenbewertungen und Quellenverweise bei Lesen und Zusammenfassen wissensbezogener Informationen aus multiplen Webseiten [Source evaluations and source references when reading and summarizing science-related information from multiple web pages]. *Unterrichtswissenschaft*, 42(1), pp. 7–23.
- Kammerer, Y., Kalbfell, E., & Gerjets, P. (2016). Is this information source commercially biased? How contradictions between web pages stimulate the consideration of source information. *Discourse Processes*, 53(5–6), pp. 430–456. doi: 10.1080/0163853x.2016.1169968
- Keck, D., Kammerer, Y., & Staruschek, E. (2015). Reading science texts online: Does source information influence the identification of contradictions within texts? *Computers & Education*, 82, pp. 442–449. doi: 10.1016/j.compedu.2014.12.005
- Kintsch, W. (1998). *Comprehension. A paradigm for cognition*. Cambridge: Cambridge University Press.
- Krell, M. (2015). Evaluating an instrument to measure mental load and mental effort using Item Response Theory. *Science Education Review Letters*, 2015, pp. 1–6.
- Kultusministerkonferenz (2012). *Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 18.10.2012)* [Educational standards in the subject German for the general qualification for university entrance (decision of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany of 18.10.2012)]. Retrieved from www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf
- Lau, J. (2006). *Guidelines on information literacy for lifelong learning*. Retrieved from archive.ifa.org/VII/s42/pub/IL-Guidelines2006.pdf
- List, A., & Alexander, P. A. (2017). Cognitive affective engagement model of multiple source use. *Educational Psychologist*, 52(3), pp. 182–199. doi: 10.1080/00461520.2017.1329014
- Locher, F. M., & Pfost, M. (2019). Erfassung des Lesevolumens in Large-Scale Studien. *Diagnostica*, 65(1), pp. 26–36. doi: 10.1026/0012-1924/a000203
- Maier, J., & Richter, T. (2013). Text belief consistency effects in the comprehension of multiple texts with conflicting information. *Cognition and Instruction*, 31(2), pp. 151–175. doi: 10.1080/07370008.2013.769997
- Mühlen, S. v. d., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). The use of source-related strategies in evaluating multiple psychology texts: A student–scientist comparison. *Reading and Writing*, 29(8), pp. 1677–1698. doi: 10.1007/s11145-015-9601-0
- Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4), pp. 429–434.

- Paul, J., Macedo-Rouet, M., Rouet, J.-F., & Stadtler, M. (2017). Why attend to source information when reading online? The perspective of ninth grade students from two different countries. *Computers & Education*, *113*, pp. 339–354. doi: 10.1016/j.compedu.2017.05.020
- Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Mahwah, NJ: Lawrence Erlbaum Associates.
- Peter, T. (2019, June 18). „Es gibt gravierende Mängel, was die Studierfähigkeit zahlreicher Abiturienten angeht“. *Leipziger Volkszeitung*. Retrieved from <https://www.lvz.de/Nachrichten/Politik/Praesident-der-Hochschulrektorenkonferenz-Es-gibt-gravierende-Maengel-was-die-Studierfaehigkeit-zahlreicher-Abiturienten-angeht>
- Primor, L., & Katzir, T. (2018). Measuring Multiple Text Integration: A Review. *Frontiers in Psychology*, *9*(2294). doi: 10.3389/fpsyg.2018.02294
- Rölke, H. (2012). The ItemBuilder: A graphical authoring system for complex item development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2012* (Vol. 2012, pp. 344–353). Chesapeake, VA: AACE.
- Rouet, J.-F., Britt, M. A., & Durik, A. M. (2017). RESOLV: Readers' representation of reading contexts and tasks. *Educational Psychologist*, *52*(3), pp. 200–215. doi: 10.1080/00461520.2017.1329015
- Rouet, J.-F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction*, *15*(1), pp. 85–106. doi: 10.1207/s1532690xci1501_3.
- Salmerón, L., Gil, L., Bråten, I., & Strømshø, H. (2010). Comprehension effects of signaling relationships between documents in search engines. *Computers in Human Behavior*, *26*(3), pp. 419–426. doi: 10.1016/j.chb.2009.11.013
- Scharrer, L., & Salmerón, L. (2016). Sourcing in the reading process [Special issue]. *Reading and Writing*, *29*(8).
- Schmalhofer, F., & Glavanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language*, *25*(3), pp. 279–294. doi: 10.1016/0749-596X(86)90002-1
- Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, *13*(2), pp. 141–156.
- Schoor, C., Hahnel, C., Artelt, C., Reimann, D., Kröhne, U., & Goldhammer, F. (2020). Entwicklung und Skalierung eines Tests zur Erfassung des Verständnisses multipler Dokumente von Studierenden [Developing and scaling a test of multiple document comprehension in university students]. *Diagnostica*. doi: 10.1026/0012-1924/a000231
- Seufert, T. (2009). Lernen mit multiplen Repräsentationen – Gestaltungs- und Verarbeitungsstrategien [Learning with multiple representations – Design and processing strategies]. In R. Plötzner, T. Leuders, & A. Wichert (Eds.), *Lernchance Computer. Strategien für das Lernen mit digitalen Medienverbänden* (pp. 45–66). Münster: Waxmann.
- Spinath, B., Stiensmeier-Pelster, J., Schöne, C., & Dickhäuser, O. (2002). *SELLMO: Skalen zur Erfassung der Lern- und Leistungsmotivation* [Learning and Achievement Motivation Scales]. Göttingen: Hogrefe.
- Stadtler, M., & Bromme, R. (2014). The content–source integration model: A taxonomic description of how readers comprehend conflicting scientific information. In D. N. Rapp & J. L. Braasch (Eds.), *Processing inaccurate information: Theoretical and ap-*

- plied perspectives from cognitive science and the educational sciences* (pp. 379–402). Cambridge, MA: MIT Press.
- Stadtler, M., Scharrer, L., Skodzik, T., & Bromme, R. (2014). Comprehending multiple documents on scientific controversies: Effects of reading goals and signaling rhetorical relationships. *Discourse Processes*, *51*(1–2), pp. 93–116. doi: 10.1080/0163853x.2013.855535
- Strømsø, H. I., Bråten, I., Britt, M. A., & Ferguson, L. E. (2013). Spontaneous sourcing among students reading multiple documents. *Cognition and Instruction*, *31*(2), pp. 176–203. doi: 10.1080/07370008.2013.769994
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), pp. 123–138. doi: 10.1007/s10648–010-9128–5
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), pp. 427–450. doi: 10.1007/bf02294627
- Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, *83*(1), pp. 73–87. doi: 10.1037/0022–0663.83.1.73



3.4

What Does It Take to Deal with Academic Literature?

Epistemic Components of Scientific Literacy

Münchow, H., Richter, T., and Schmid, S.

Abstract

The skills required for understanding and evaluating academic literature include a broad repertoire of different reading strategies, which are rarely explicitly taught and go beyond classical learning strategies that foster learning from expository texts. This chapter proposes a taxonomy of strategies for reading academic literature, which distinguishes between two different processing goals (receptive vs. epistemic) and processing modes (systematic vs. heuristic). Recent research on epistemic-systematic reading strategies, diagnostic instruments to assess these strategies, and training interventions to foster these strategies is described in more detail. Finally, the chapter provides an outlook on further research that includes epistemic-heuristic reading strategies as another key component of scientific literacy.

Keywords

Argument comprehension, argument evaluation, epistemic learning strategies, receptive learning strategies, scientific literacy

Funding

Preparation of this paper was supported by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF); grant 01PK15009A and grant 01PK15009B.

1 What Does It Take to Deal with Academic Primary Literature? More Than Deep-Level Learning Strategies

University students are often advised to organize the information they are reading, to relate it to relevant prior knowledge, and to reread passages they find difficult to understand (e.g., Pintrich 2004). Other suggestions to improve learning from text include the use of practice tests (Karpicke and Roediger 2010) or the distribution of rereading across time (Rawson and Kintsch 2005). This advice is based on a broad body of research on learning strategies, reading strategies, and self-regulated learning in cognitive and educational psychology, rooted in theories of human cognitive architecture (Weinstein and Mayer 1986) and supported by a wealth of correlational studies and training experiments (for reviews, see Dunlosky et al. 2013; Richardson et al. 2012). Certainly, the learning strategies that have been in the focus of cognitive and educational research for many years can help students to deeply comprehend expository texts, remember the information conveyed by these texts, and retrieve it from long-term memory at a later test. However, there is doubt whether these strategies suffice for appropriately dealing with other types of academic literature and in other types of reading situations. Academic primary literature, in particular, might call for the application of strategies that have rarely been studied in previous research.

In the first part of this chapter, we briefly sketch *a taxonomy of strategies* that goes beyond traditional models of reading and learning strategies (e.g., Pintrich 2004; Weinstein and Mayer 1986) and that covers a broader range of reading situations that university students are typically confronted with. We seek to approximate a more comprehensive conceptualization of reading strategies that together form the core of scientific literacy. Scientific literacy can be defined as “the ability to understand and critically evaluate scientific content to achieve one’s goals” (Britt et al. 2014, p. 104). In line with this definition, we distinguish two generic reading goals, which we call *receptive and epistemic goals*, and two modes of understanding and evaluating scientific content, the *heuristic and the systematic mode*. Receptive reading goals involve comprehending and memorizing information (e.g., for recalling this information in a later test); epistemic reading goals

involve the use of texts for the acquisition of knowledge the reader regards as plausible or true (Richter 2003). The *epistemic reading strategies* that are functional for accomplishing epistemic reading goals are novel in the sense that they are not covered by classical taxonomies of reading learning strategies (e.g., Weinstein and Mayer 1986) and have only recently come to the fore of educational psychology research (Barzilai and Zohar 2014; Richter and Schmid 2010). In the second part of the chapter, we focus on systematic epistemic reading strategies. More specifically, we give an overview of recent research from our lab on the characteristics of systematic epistemic reading strategies, how these strategies can be assessed, and how they can be taught to university students through systematic, computer-based trainings.

2 A Taxonomy of Reading Strategies for Academic Literature

Scientific content is mainly communicated through written texts. These texts are typically written by scientists to be read by scientists (Goldman and Bisanz 2002). Yet, university students of almost all disciplines are required to read academic literature as part of their studies. To illustrate the scope of the learning strategies required to comprehend academic literature, it is instructive to consider the way scientists themselves read scientific texts. The extant studies (Bazerman 1985; Berkenkotter and Huckin 1995) indicate that scientists routinely employ a broad repertoire of reading strategies, which include both heuristic and epistemic strategies. The physicists interviewed by Bazerman (1985), for example, distinguished between core reading and peripheral reading. In their peripheral reading, they scanned texts for particular words, skipped sections, and evaluated publications by their authors. Remarkably, projected onto classical taxonomies of learning strategies, these reading behaviors routinely exhibited by experts in the domain of physics would have to be regarded as superficial processing, although they apparently help the physicists to achieve certain goals: for example, to find specific information in a text or to select those texts for a closer reading that seem to be most informative or trustworthy.

Our taxonomy covers the variety of reading situations and goals that scientists and students alike are confronted with by posing two generic reading goals and two modes of processing scientific information. First, we distinguish between two types of goals: receptive vs. epistemic goals. Readers with a receptive goal strive to learn facts, understand the text contents, or find specific information, whereas readers with an epistemic goal strive to gain an adequate picture of the state-of-

affairs described in the text or develop an own standpoint on the issues discussed in the text (Richter 2003, 2011; Richter and Schmid 2010). Receptive reading goals can be fully described within the classical information processing framework of cognitive psychology (Neisser 1967); learning strategies that help achieving receptive learning goals improve cognitive processes such as encoding, manipulation, storage, and retrieval of information. However, how and to what extent the processed information corresponds to state-of-affairs in the world, whether it is true or whether there are good reason to assume that it is true, is irrelevant for the attainment of receptive reading goals. For example, students studying for an exam need to comprehend and memorize information in expository texts, but it is often irrelevant to their study goal whether the theories described in the text are actually valid. By contrast, for epistemic reading goals, these criteria are essential. Thus, epistemic reading goals involve the acquisition of knowledge in a classical (philosophical) sense, that is, the acquisition of true and justified beliefs (e.g., Ichikawa and Steup 2018). For example, a student who aims at identifying the most powerful theory in a given field of study would follow an epistemic reading goal.

Second, we distinguish between systematic (deep-level) or heuristic (surface-level) strategies (following the distinction made in two-process models of information processing, e.g., Petty and Wegner 1999). Systematic strategies involve controlled processes, are cognitively demanding, and, if successful, lead to deeper understanding of a scientific issue or domain. By contrast, heuristic strategies can be applied fast and demand less cognitive resources. If successful, heuristic strategies often result in a specific decision, for example, that a piece of information matches the answer of a question, that a statement is implausible, that a document is trustworthy, or that it is worthwhile to read a text more thoroughly. The distinction between systematic and heuristic strategies does not imply that systematic strategies lead to better results than heuristic strategies. Rather, it depends on the processing goal and other conditions (such as the available time, reader's expertise, and the information provided by the text) whether a systematic or heuristic strategy is appropriate.

Combining the dimensions epistemic-receptive and systematic-heuristic yields a 2x2 table with four categories of strategies (Figure 1). The four categories are described next with examples.

Processing Mode	Processing Goal	
	Receptive	Epistemic
Systematic	e.g., organisation	e.g., evaluation of the consistency of arguments
Heuristic	e.g., scanning for certain information	e.g., using source information

Figure 1 Overview of taxonomy of reading and learning strategies required to deal with academic literature

2.1 Receptive-Systematic Strategies

The strategies in this category serve to enrich or structure information to facilitate later recall. Examples are classical learning strategies such as rehearsal, organization and elaboration, as described in the educational and cognitive research on reading and learning strategies (e.g., Dunlosky et al. 2013; Pintrich 2004).

2.2 Receptive-Heuristic Strategies

The strategies in this category serve to gain a first impression of the text content. Readers can rely on genre knowledge about the types of information texts in a particular genre typically provide. Once a reader has selected a particular text for further scrutiny, he or she can turn to receptive-heuristic strategies such as *skimming* (cursory reading to extract the gist) and *scanning* (cursory reading to find specific information). These strategies are particularly important in the phase of literature research and require generic knowledge of canonic text structures (Dillon 1991).

2.3 Epistemic-Systematic Strategies

The strategies in this category serve to validate the argumentation of a text. In some models of learning strategies, they are covered by the construct *critical evaluation* (e.g., Pintrich 2004). At a more specific level, it is possible to distinguish between *consistency checking* and *knowledge-based validation* (Richter and Schmid 2010). These strategies require the reader to identify the functional components

of arguments such as *claim* and *ground* (reasons), to evaluate the acceptability of the reasons, and to evaluate the internal consistency of the argument, i.e., the relevance and sufficiency of the reasons (Larson, Britt and Kurby 2009; Toulmin 1958). University students often fail to identify the argumentative function of text passages (e.g., Norris et al. 2003).

2.4 Epistemic-Heuristic Strategies

The strategies in this category serve to gain a quick preliminary evaluation of the credibility of the text. Examples include the use of source information (e.g., publication outlet and funding, Zimmerman et al. 2001), fast judgments of the plausibility of claims (Voss et al. 1993), or the selective processing of belief-consistent information (Maier and Richter 2013; Richter and Maier 2017). Such strategies are particularly important when readers lack domain-specific content knowledge or the cognitive or motivational resources necessary for deep processing. Contrary to scientists, students often neglect source information when making epistemic judgments (Zimmerman et al. 2001), which may hamper their comprehension of scientific information (e.g., Strømsø, Bråten and Britt 2010).

University students need to possess a broad knowledge of strategies in all four categories for competently dealing with academic literature. However, they also need to know when and for what purpose they can use a particular strategy (conditional knowledge, Lorch et al. 1993). Skimming and scanning, for example, are particularly important in the phase of literature research to find out whether a particular publication is relevant for the question at hand. Only if this condition is met, the reader should turn to systematic strategies. To give another example, readers following an epistemic reading goal might not be able to adequately judge the arguments in a text as they lack pertinent prior knowledge. In that case, the epistemic reading goal might be served better by making a heuristic judgement about the credibility of the source, even if the text presents the scientific issue in a way that it seems easy to comprehend (Scharrer et al. 2014).

3 Epistemic-Systematic Reading Strategies: Assessment, Training, and Relevance for Studying at University

In the remainder of this chapter, we present an overview of research from our own lab on epistemic reading strategies, i.e., the strategies required for comprehending and evaluating the arguments presented in scientific texts. Epistemic reading strategies go beyond a receptive elaboration of the texts' contents and enable students to acquire knowledge about scientific issues as opposed to identifying and memorizing information. Such reading strategies are usually not explicitly taught in school, which contributes to the problems many first-year students encounter when reading academic primary literature. Deficits in epistemic strategies were revealed, for example, in student-scientist comparisons that examined how psychology students and scientists (advanced doctoral students and postdocs) comprehended and evaluated arguments (von der Mühlen et al. 2016a, 2016b). In these studies, scientists were superior in decoding the functional structure of informal arguments, in identifying implausible arguments, and in recognising errors in argumentation. These findings underline the need for diagnostic instruments for the assessment of epistemic-systematic strategies and for training interventions that foster these strategies in university students. In a number of studies to be described next we developed and evaluated such assessments and trainings for undergraduates in the social sciences.

3.1 Assessment of Epistemic-Systematic Reading Strategies

For assessing students' epistemic-systematic reading strategies, we developed two computer-based diagnostic instruments: the Argument Structure Test (AST, Münchow et al. 2020) and the Argument Judgement Test (AJT, Münchow et al. 2019).

3.1.1 Argument Structure Test

The Argument Structure Test assesses students' ability to identify the structural components of informal arguments, which may be considered the building blocks of academic literature. The test consists of eight short informal arguments ($M = 104$ words, $SD = 24$ words) taken from typical psychological texts. The arguments are composed of argument components according to Toulmin (1958). According to this model, arguments consist of up to five functionally different components: The claim or the statement that is being argued for, one or several (empirical, theoretic-

tical, or practical) reasons that support the claim, a warrant that states why the reason(s) should support the claim, a backing that justifies the warrant empirically or theoretically, and a rebuttal that limits the validity of the claim (e.g., by referring to exceptions). In empirical sciences, the claims found in scientific texts are most often theoretical in nature (e.g., an explanatory assumption) and the reasons are findings from empirical studies. This principle is also followed in the arguments used in the Argument Structure Test. Respondents' task is to assign the five functional arguments components to one of the sentences of each of the eight arguments. The arguments used in the Argument Structure Test differ with regard to several characteristics that are known to affect argument comprehension. First, the position of claim and reasons is varied. In half of the arguments, the claim is located in the first sentence of the argument (claim-first arguments), whereas the other half of the arguments start with a reason (reason-first arguments). Claim-first arguments follow the canonical order and are easier to comprehend as the claim is the key to building up a mental representation of the argument (Britt and Larson 2003). Moreover, argument complexity is varied systematically. Two arguments are simple arguments that contain only three of the five argument components distinguished by Toulmin, whereas the remaining six arguments contain all five argument components (complex). The two simple arguments are claim-first arguments.

Respondents first read an argument as a continuous text (Figure 2A). Afterwards they are shown the same argument again and are asked to assign the sentences of the argument to one of the five argument components (claim, reason, warrant, backing, and rebuttal) via a dropdown menu (Figure 2B). Students usually need 20 to 30 minutes to complete all items of the Argument Structure Test. The total number of accurately assigned argument components serves as a score of the reader's ability to decode the functional structure of informal arguments.

Psychometric properties of the Argument Structure Test were evaluated in a study with a convenience sample of 225 psychology undergraduates and teacher students. In this study, the internal consistency (*Cronbach's alpha*) of the Argument Structure Test reached .76, with a wide range of item difficulties ($M = .69$, $SD = .16$). The items of the Argument Structure Test showed a good fit to the Rasch model (1-PL model, Andersen Likelihood-Ratio test with a mean-split of the sample: $\chi^2[df = 38, N = 225] = 46.81, p = .130$), with no indication of interdependencies between items within a specific argument.

To examine the construct validity of the Argument Structure Test, we estimated an explanatory item response model (LLTM, Fischer 1974) to predict the test's item difficulties through theoretically relevant item characteristics (order of argument components and arguments complexity). As expected, claim-first arguments and less complex arguments were easier to decode. Finally, the observed item diffi-

culties based on the Rasch model could be predicted well through item difficulties estimated with the LLTM ($R^2 = .82$, Figure 3), which is strong evidence for the construct validity of the instrument.

A

Please read the text carefully before clicking the CONTINUE button.

Self-control can predict success at school and should be trained as early as possible. In a longitudinal study with 653 children, Mischel and Shoda (1988) investigated how the willingness to postpone a reward affects the development of school children. The authors found that children who postponed a reward (e.g. a biscuit) at the age of four or five if another reward (two biscuits) was promised had better cognitive and social skills ten years later than children who preferred an immediate reward. Success at school plays a central role for further professional success. High school graduates with very good grades, for example, often have better student-teacher relationships during their studies, less difficulties and stress, and a more stable course of studies (Bargel, 2002). Of course, in addition to successful self-control, there are many other factors that are responsible for a child's school development..

CONTINUE**B**

Now you see the text again segmented into its different components. Your task is to correctly identify the elements of the argument structure of this text.

Assign the claim/conclusion, reason, warrant, backing of the warrant and rebuttal to the correct number. If you are unsure, please select "I don't know". If you are of the opinion that a component does not occur, please select "does not occur".

1. Self-control can predict success at school and should be trained as early as possible.
2. Mischel and Shoda (1988) investigated how the willingness to postpone a reward affects the development of school children. The authors found that children who postponed a reward (e.g. a biscuit) at the age of four or five if another reward (two biscuits) was promised had better cognitive and social skills ten years later than children who preferred an immediate reward.
3. Success at school plays a central role for further professional success.
4. High school graduates with very good grades, for example, often have better student-teacher relationships during their studies, less difficulties and stress, and a more stable course of studies (Bargel, 2002).
5. Of course, in addition to successful self-control, there are many other factors that are responsible for a child's school development.

Which number corresponds to the claim/conclusion?

An argumentative claim/conclusion is a controversial thesis which an author tries to convince readers of by citing theoretical or practical (e.g., ethical) reasons or empirical evidence.

CONTINUE

Figure 2 Example item for the Argument Structure Test. (A) Argument presented as a continuous text. (B) Argument separated by sentences (translated from German). (Figure adapted from Münchow et al. 2020, p. 3)

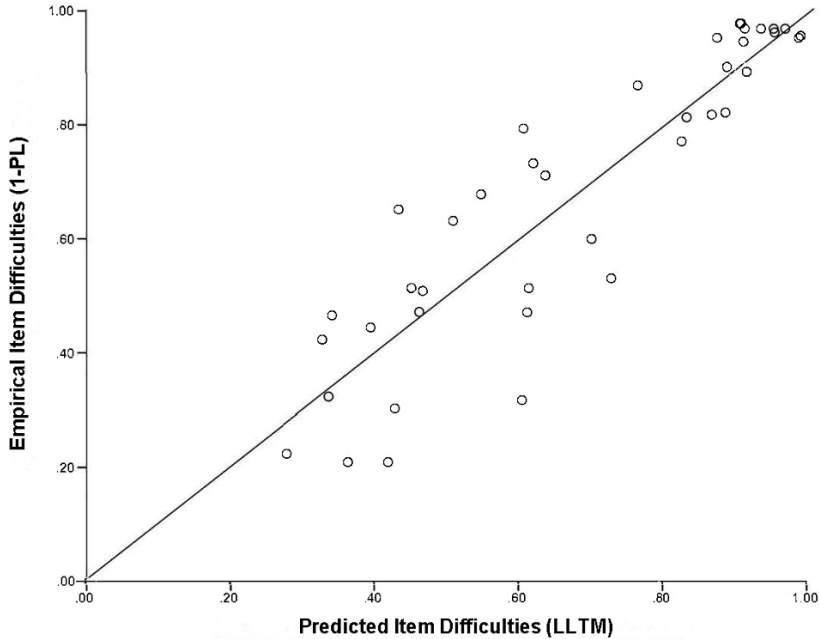


Figure 3 Scatterplot of predicted and empirical item difficulties for the Argument Structure Test (Figure adapted from Münchow et al. 2020, p. 8)

Evidence for the criterial validity of the test was obtained through correlations of the test scores with criterial performance measures as well as the students' epistemological beliefs (scales Structure and Variability of the Connotative Aspects of Epistemic Beliefs Questionnaire, CAEB; Stahl and Bromme 2007). The test scores of the Argument Structure Test were moderately and significantly ($p < .05$) correlated with verbal intelligence ($r = .40$) assessed with the subtests sentence completion, analogies and commonalities of the Intelligence Structure Test (I-S-T 2000 R, Amthauer, Brocke, Liepmann and Beauducel 2001) and students' Grade Point Average in the school-leaving certificate (Abitur; $r = .17$). Moreover, students who scored higher in the Argument Structure Test were more likely to see knowledge in psychology as structured but changeable, as indicated by significant correlations with the scales Structure ($r = .20$) and Variability ($r = -.33$, reverse scored) of the CAEB. In sum, the Argument Structure Test is a reliable and valid instrument for assessing the ability to comprehend arguments in scientific texts.

3.1.2 Argument Judgement Test

The Argument Judgement Test assesses students' abilities to accurately judge the plausibility of informal arguments and to identify common argumentation fallacies. The test consists of two parts. In Part 1 of the Argument Judgement Test, readers are presented two short expository texts about smoking behavior (550 words) and objective self-awareness (404 words). Each text consists of 15 short informal arguments containing a claim and one or several reasons. Twenty out of 30 arguments are plausible, that is, these arguments contain strong and internally consistent reasons that support the claim. The remaining ten arguments contain one of five common argumentation fallacies (i.e., contradiction, false dichotomy, wrong example, circular reasoning, overgeneralization; Dauer 1989), resulting in poor arguments. The readers' task in Part 1 of the Argument Judgment Test is to evaluate whether the presented arguments are plausible or implausible by pressing a corresponding key on the keyboard. The number of correctly judged arguments serves as test score. In Part 2, the arguments that the respondents judged as implausible in Part 1 are presented again, and respondents are asked to assign each of these arguments to one of the five common argumentation fallacies via a dropdown menu. Each fallacy is briefly explained on screen. Participants can also select the answer options *I don't know*, *I was wrong*, *there is no error*, or *None of the above-mentioned errors, but ...*, where they can enter text in a text box for the last option. The number of correctly assigned arguments is used as a test score in Part 2. Furthermore, a combined score of the responses in Part 1 and Part 2 can be formed, which is a measure of the ability to evaluate the plausibility of informal arguments. An example item of the Argument Judgement Test is shown in Figure 4.

The psychometric properties of the Argument Judgment Test were explored in a study (Münchow et al. 2019) based on the same convenience sample of 225 psychology and teacher students used for examining the Argument Structure Test. In this study, Part 1 turned out to be relatively easy (item difficulty: $M = .74$, $SD = .10$), whereas Part 2 was relatively difficult (item difficulty: $M = .36$; $SD = .22$). However, a combined score of Part 1 and Part 2 responses lead to a wide distribution of item difficulties. The items of the combined scale showed a good fit with the Rasch model (1-PL model, Andersen LR-test based on a mean-split of the sample: $\chi^2[df = 25, N = 225] = 27.53, p = .330$). The internal consistency for the combined score was acceptable (with a WLE reliability coefficient of .63), and the stability (test-retest reliability) within an interval of 13 months reached .60 in an independent sample of 22 psychology students.

A The construct of inherited nicotine sensitivity seems to play a central role here. This construct refers to the fact that some people react more strongly to nicotine because they are more sensitive to nicotine.

Press the P key if the sentence seems plausible, or the Q key if the reasoning does not seem plausible.

“Q”
Implausible
“P”
Plausible

B The construct of inherited nicotine sensitivity seems to play a central role here. This construct refers to the fact that some people react more strongly to nicotine because they are more sensitive to nicotine.

You have declared the above sentence implausible. Please select the argumentation error you think is involved.

- Circular Reasoning
[The attempt to prove the correctness of a premise with the help of a (logical) conclusion drawn from this premise, The premises shall prove the conclusion and at the same time the conclusion shall prove the premise.]
- Classical False Conclusion - Overgeneralization
[A premise is followed by a false, hasty conclusion by generalizing or overrating results.]
- Classical False Conclusion – Contradiction
[A premise is followed by an incompatible conclusion.]
- Wrong Example
[A false or inappropriate example is cited as evidence for an allegation.]
- Wrong Dichotomy
[A contradiction is suggested, but it's not really a contradiction.]
- I don't know.
- I was wrong, there is no error.
- None of the above mentioned errors, but

NEXT

Figure 4 Example item for (A) Part 1 and (B) Part 2 of the Argument Judgement Test (translated from German) (Figure adapted from Münchow et al. 2019, p. 6)

Construct validity of the Argument Judgment Test was examined by estimating an explanatory item response model revealing that implausible arguments were more difficult to detect than plausible arguments. In addition, we estimated linear mixed models (items \times participants) with the students' response times in Part 1 of the Argument Structure Test as the dependent variable. The main idea behind these analyses was that argument evaluation is a rational, effortful activity, which should lead to longer response times in implausible arguments, if these arguments were indeed recognized as implausible. Indeed, there was a significant interaction effect ($p < .05$) of response accuracy and argument plausibility that followed the expected pattern (Figure 5). These results provide some evidence that effortful processing is needed for accurately evaluating implausible arguments, whereas plausible arguments can be evaluated more efficiently.

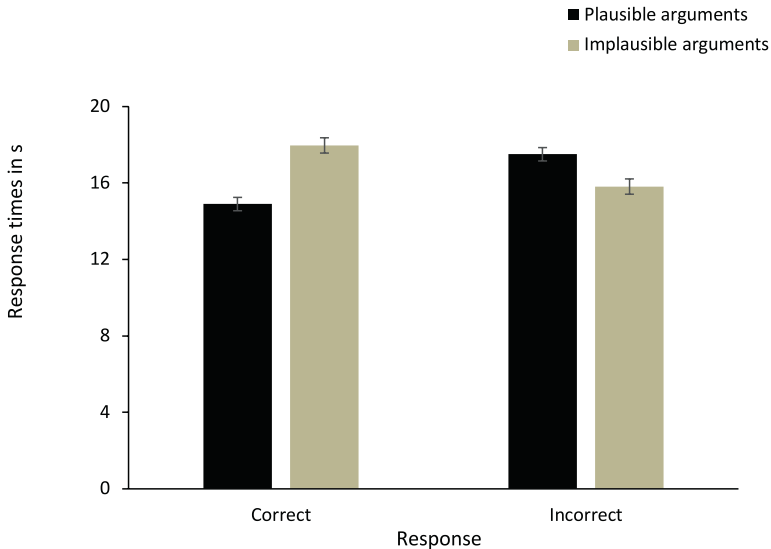


Figure 5 Response times (in seconds) in Part I of the Argument Judgement Test for plausible versus implausible arguments and response accuracy (Figure adapted from Münchow et al. 2019, p. 14)

With regard to criterial validity, test scores were moderately and significantly correlated ($r = .36$) with verbal intelligence (assessed with the verbal subtests of the I-S-T 2000 R, Amthauer et al. 2001) and the students' current average grade as well as the student's epistemological beliefs (again measured with the CAEB). Moreover, AJT test scores were significantly associated with students' academic success measured via the students' current grade average at university, even if verbal intelligence and the students' Grade Point Average from school leaving certificates were controlled for. The increment of explained variance was 20 %.

3.2 Training of Epistemic-Systematic Reading Strategies

The studies conducted with the Argument Structure Test and the Argument Judgement Test suggest that large individual differences exist in the ability to comprehend and evaluate informal arguments (Münchow et al. 2019, 2020). Moreover, first year students show deficits in epistemic-systematic strategies (e.g., von der Mühlen 2016a), presumably because these strategies are not explicitly taught in

school, where scientific knowledge is often presented as a “monolith of facts” (Osborne 2010, p. 464) rather than the product of rational, argument-based discourse. Thus, there is a need for effective trainings to foster epistemic reading strategies in university students. We addressed this need by developing and evaluating two computerized trainings, an *Argument Structure Training* to enhance students’ argument decoding skills and an *Argument Judgment Training* to foster students’ skills of evaluating the internal consistency and validity of the argumentation of a text. Both trainings convey conceptual knowledge about arguments (their functional components or normative criteria for their evaluation respectively), which is presented via illustrated texts, audio examples, and short video clips. Various exercises during and after the theoretical parts give participants the opportunity to apply and practice the content they have learned. Participants receive direct feedback after each task and can either redo that specific task, return to the corresponding theory block, or continue training. The trainings last about 45–60 minutes, but there is no limit to the training time. The evaluation of the trainings was based on the Argument Structure Test and the Argument Judgement Test, respectively.

3.2.1 Argument Structure Training

The argument structure training imparts strategies for evaluating scientific arguments by training the identification and allocation of functional argument components. The theoretical input focuses on the use and purpose of informal arguments, the Toulmin (1958) model of argumentation, and linguistic connectors and key words that help to correctly identify certain argument components. Figure 6 shows example pages of the Argument Structure Training.

The effectiveness of the argument structure training was evaluated in two experimental pre-post-test studies with a follow-up test four weeks after the training. Both studies employed an active control group that received a computerized speed reading training that involved reading but did not train epistemic-systematic reading strategies in any way. Participants of the first training experiment were 53 psychology students at the beginning of their studies. Students in the training condition outperformed participants in the control condition in their ability to identify and assign less typical argument components (i.e., warrants) and to correctly identify argument components in arguments with a less typical structure, i.e., reason-first arguments (medium-sized effects significant at $p < .05$). Moreover, the training intervention was especially effective for students that, according to their grades, were more successful in their studies (von der Mühlen et al. 2018). However, there were no differences between the two conditions in their argument structure decoding skills at follow-up measures, which indicates that the training

effects did not remain stable over a period of four weeks. We therefore conducted a second training experiment to evaluate effectiveness of the Argument Structure Training plus a 15-minute booster training session in the week before the follow-up tests. Analyses of the data from this study are still in progress.

3.2.2 Argument Judgment Training

The argument judgment training teaches strategies for the normatively appropriate evaluation of arguments, especially strategies for evaluating the relevance and completeness of reasons for the justification of an argument's claim, and trains students to recognize typical errors in argumentation, such as circular reasoning or overgeneralization.

The Argument Judgment Training was also evaluated in an pre-post-test experimental design with an active control group (speed-reading training) and a follow-up after four weeks (as yet unpublished study). Psychology students and teacher students participated in this study. Similarly to the results for the Argument Structure Training, participants performed better in the training condition than in the control condition at post-test, but the effect disappeared in the post-test after four weeks.

In sum, both trainings developed to foster epistemic-systematic reading strategies in university students produced immediate effects on the trained skills, but the effects were not stable over time. Moreover, evidence for transfer effects, for example on study performance, is still lacking. We are planning future studies to address these issues. In these studies, we plan to add additional instructional measures such as practice tests (Greving and Richter 2018) and interleaved presentation of content (Brunmair and Richter 2019) to the training that promise to foster long-term and transfer effects.

A

Why should you train how to deal with arguments?

Because a good handling of arguments is a basic requirement for a successful study! Arguments are fundamental components of scientific texts and you will often encounter them during your studies. If you know how arguments work, you can better understand and evaluate scientific texts, uncover their strengths and weaknesses. Especially for beginners such a focus on text structure offers the chance to compensate the lack of experience with scientific literature.

The aim of this training is to enable you to answer the following questions at the end:



1. What is the structure of an argument?
2. What components does an argument have?
3. How can you identify them?

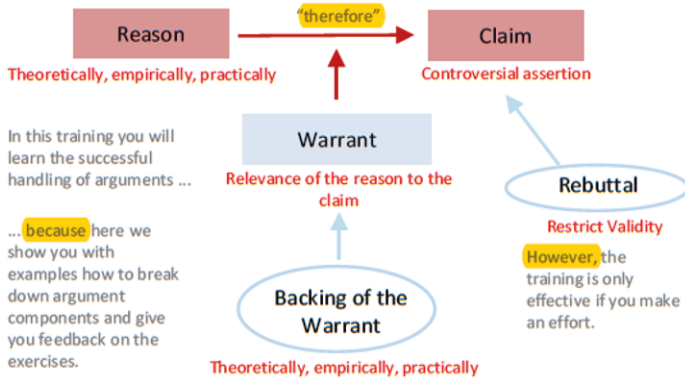
BACK

CONTINUE

B

Because arguments play an important role in science ...

... it is important to complete this training.



BACK

CONTINUE

Figure 6 Example pages of the Argument Structure Training showing (A) the goals of the training and (B) the Toulmin model of argumentation (1958) and typical linguistic connectors

4 Conclusion and Outlook

In this chapter, we outlined a taxonomy of reading strategies that differentiates between epistemic-systematic strategies, epistemic-heuristic strategies, receptive-systematic strategies, and receptive-systematic strategies, and described tests and trainings that target epistemic-systematic strategies. Epistemic-systematic strategies are particularly relevant for adequately dealing with scientific literature, but have been neglected in previous research. Other tests and trainings from our lab also cover epistemic-heuristic strategies. The tests for the assessment of these strategies (e.g., the Credibility Judgment Test; von der Mühlen et al. 2016a) involve different sets of texts, which are presented for a limited time to prevent students from engaging in systematic text processing. Thus, unlike other researchers (e.g., in the field of sourcing research; Wineburg 1991; Brante and Strømsø 2018; see also Schoor et al. in this volume), we followed an approach of employing different types text materials and procedures for the assessment of heuristic and systematic components of scientific literacy. An advantage of this approach is that it allows analyzing relationships between these components. The correlation between epistemic-systematic competencies and epistemic-heuristic competencies was significant, but moderate, $r = .58, p < .05$ (von der Mühlen et al. 2016). It thus seems that scientific literacy is composed of distinguishable facets (Britt et al. 2014).

The Argument Structure Test and the Argument Judgement Test, and the corresponding trainings, were constructed as research tools. Nevertheless, the findings based on these tests and trainings indicate that they could also be used for practical purposes. The reliabilities of the tests are too low to warrant responsible individual-level selection decisions, but high enough to detect deficits on the group-level or to evaluate university courses. With regard to the trainings, our studies demonstrate that even short-term interventions can improve students' epistemic-systematic competencies. From a practical perspective, more extensive interventions are needed to ensure sustainable effects. For example, our trainings could be used as a starting point of regular reading courses, to familiarize students with the structural components of informal arguments and common argumentation fallacies. Subsequent sessions could then be devoted to the analysis and discussion of the way arguments are typically laid out in concrete exemplars of academic literature from the respective domain, to forge links between content knowledge, genre knowledge, and knowledge about argumentation. Such courses would not only provide students with disciplinary content knowledge, but also enable them to understand, and evaluate, the arguments on which this knowledge rests.

References

- Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *I-S-T 2000 R – Intelligenz-Struktur-Test 2000 R* [Intelligence Structure Test 2000 R]. Göttingen, Germany: Hogrefe.
- Barzilai, S., & Zohar, A. (2014). Reconsidering personal epistemology as metacognition: A multifaceted approach to the analysis of epistemic thinking. *Educational Psychologist, 49*, 13–35.
- Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written Communication, 2*, 3–23.
- Berkenkotter, C., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognition/ culture/ power*. Hillsdale, NJ: Erlbaum.
- Brante, E. W., & Strømsø, H. I. (2018). Sourcing in text comprehension: A review of interventions targeting sourcing skills. *Educational Psychology Review, 30*, 773–799.
- Britt, M. A., & Larson, A. A. (2003). Constructing representations of arguments. *Journal of Memory and Language, 48*, 794–810.
- Britt, M. A., Richter, T., & Rouet, J. F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist, 49*, 104–122.
- Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*. doi: 10.1037/bul0000209
- Dauer, F. W. (1989). *Critical thinking: An introduction to reasoning*. New York, NY: Oxford University Press.
- Dillon, A. (1991). Reader's models of text structures: The case of academic articles. *International Journal of Man-Machine Studies, 35*, 913–925.
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58.
- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to the theory of psychological testing]. Bern: Huber.
- Goldman, S. R., & Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J.A. León, & A.C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 417–436). Mahwah, NJ: Erlbaum.
- Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology, 9*:2412. <https://doi.org/10.3389/fpsyg.2018.02412>
- Ichikawa, J. J., & Steup, M. (2018). The analysis of knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* [Online Document]. Retrieved from <https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/>
- Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition, 38*, 116–124.
- Larson, A. A., Britt, M. A., & Kurby, C. A. (2009). Improving students' evaluation of informal arguments. *Journal of Experimental Education, 77*, 339–366.

- Lorch, R. F., Lorch, E. P., & Klusewitz, M. A. (1993). College students' conditional knowledge about reading. *Journal of Educational Psychology, 85*, 239–252. <http://dx.doi.org/10.1037/0022-0663.85.2.239>
- Maier, J., & Richter, T. (2013). Text-belief consistency effects in the comprehension of multiple texts with conflicting information. *Cognition and Instruction, 31*, 151–175.
- Münchow, H., Richter, T., von der Mühlen, S., & Schmid, S. (2019). The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, nomological network and relevance for academic success at the university. *British Journal of Educational Psychology, 89*, 501–523. <https://doi.org/10.1111/bjep.12298>
- Münchow, H., Richter, T., von der Mühlen, S., Schmid, S., Bruns, K. & Berthold, K. (2020). Verstehen von Argumenten in wissenschaftlichen Texten: Reliabilität und Validität des Argumentstrukturtests (AST) [Comprehension of arguments in scientific tests: Reliability and validity of the argument structure test]. *Diagnostica*.
- Neisser, U. (1967). *Cognitive Psychology*. New York, NY: Psychology Press.
- Norris, S. P., Phillips, L.M., & Korpan, C.A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty. *Public Understanding of Science, 12*, 123–145.
- Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*(5977), 463–466. <https://dx.doi.org/10.1126/science.1183944>
- Petty, R. E., & Wegener, D. T. (1999). The elaboration likelihood model: Current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 41–72). New York: Guilford Press.
- Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*, 385–407.
- Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*, 70–80. <https://doi.org/10.1037/0022-0663.97.1.70>
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 352–387.
- Richter, T. (2003). *Epistemologische Einschätzungen beim Textverstehen* [Epistemic validation in text comprehension]. Lengerich: Pabst.
- Richter, T. (2011). Cognitive flexibility and epistemic validation in learning from multiple texts. In J. Elen, E. Stahl, R. Bromme, & G. Clarebout (Eds.), *Links between beliefs and cognitive flexibility* (pp. 125–140). Berlin: Springer.
- Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A Two-step Model of Validation. *Educational Psychologist, 52*, 148–166. <http://dx.doi.org/10.1080/00461520.2017.1322968>
- Richter, T., & Schmid, S. (2010). Epistemological beliefs and epistemic strategies in self-regulated learning. *Metacognition and Learning, 5*, 47–65.
- Scharrer, L., Stadler, M., & Bromme, R. (2014). You'd better ask an expert: Mitigating the comprehensibility effect on laypeople's decisions about science-based knowledge claims. *Applied Cognitive Psychology, 28*, 465–471.
- Stahl, E., & Bromme, R. (2007). The CAEB: An instrument for measuring connotative aspects of epistemological beliefs. *Learning and Instruction, 17*, 773–785.

- Strømsø, H. I., Bråten, I., & Britt, M. A. (2010). Reading multiple texts about climate change: The relationship between memory for sources and text comprehension. *Learning and Instruction, 20*, 192–204.
- Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.
- von der Mühlen, S., Richter, T., Schmid, S., & Berthold, K. (2018). How to improve argumentation comprehension in university students: Experimental test of a training approach. *Instructional Science, 47*, 215–237.
- von der Mühlen, S., Richter, T., Schmid, S., Schmidt, L. M., & Berthold, K. (2016a). The use of source-related strategies in evaluating multiple psychology texts: A student-scientist comparison. *Reading and Writing, 8*, 1677–1698.
- von der Mühlen, S., Richter, T., Schmid, S., Schmidt, L. M., & Berthold, K. (2016b). Judging the plausibility of arguments in scientific texts: A student-scientist comparison. *Thinking & Reasoning, 22*, 221–246.
- Voss, J. F., Fincher-Kiefer, R., Wiley, J., & Silfies, L. N. (1993). On the processing of arguments. *Argumentation, 7*, 165–181.
- Weinstein, C. E., & Mayer, R. E. (1986). The teaching of learning strategies. In M.C. Wittrock (Ed.), *Handbook of research in teaching* (pp. 315–327). New York, NY: Macmillan.
- Wineburg, S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83*, 73–87.
- Zimmerman, C., Bisanz, G. L., Bisanz, J., Klein, J. S., & Klein, P. (2001). Science at the supermarket: A comparison of what appears in the popular press, expert's advice to readers, and what students want to know. *Public Understanding of Science, 10*, 37–58.



3.5 Measuring Scientific Reasoning Competencies

Multiple Aspects of Validity

Krüger, D., Hartmann, S., Nordmeier, V., and
Upmeier zu Belzen, A.

Abstract

In this chapter, we investigate multiple aspects of validity of test score interpretations from a scientific reasoning competence test, as well as aspects of reliability. Scientific reasoning competencies are defined as the disposition to solve scientific problems in certain situations by conducting scientific investigations or using scientific models. For the purpose of measurement, the first phase of our project focused on the construction of a paper-pencil assessment instrument – the *KoWADiS* competence test – for the longitudinal assessment of pre-service science teachers' scientific reasoning competencies over the course of academic studies. In the second phase of our project, we investigated the reliability of the test scores and the validity of their interpretations. We used a multimethod approach, addressing several sources of validity evidence. Overall, the results are coherent and support the validity assumptions to a satisfactory degree. The long-term goal is the use of this test to provide empirically sound suggestions for pre-service science teacher education at university level.

The electronic edition of this chapter has been revised: The university name for the authors S. Hartmann and A. Upmeier zu Belzen has been corrected. A Correction is available at https://doi.org/10.1007/978-3-658-27886-1_21

© Springer Fachmedien Wiesbaden GmbH, part of Springer Nature 2020,
corrected publication 2021

O. Zlatkin-Troitschanskaia et al. (Hrsg.), *Student Learning in German
Higher Education* https://doi.org/10.1007/978-3-658-27886-1_13

Keywords

Scientific reasoning competencies, teacher education, validity

1 Introduction

Scientific reasoning competencies are a set of acquired cognitive dispositions that individuals such as scientists, teachers, or students use to solve scientific problems systematically. The assessment of cognitive dispositions is possible by measuring manifest behavior, i.e. performance (Koeppen et al. 2008). Scientific reasoning competencies are performed by applying skills like planning, conducting, and evaluating scientific investigations, or using scientific models (Fischer et al. 2014). These inquiry processes allow to gain new insights into scientific phenomena, utilizing research questions, theories and hypotheses as typical aspects of the hypothetico-deductive approach of the empirical sciences (Popper 2004). Competence tests as construct-related measurements are needed to assess what students are able to do as a result of their learning (Blömeke et al. 2015; Osborne 2013). Competence tests indicate students' performance based on reliably and validly interpretable criterion-related measures. As sources of evidence for validity we investigated test content, response processes, internal structure, and relations to other variables (AERA et al. 2014). These criteria serve as sources of evidence and were applied systematically to test our instrument, the *Ko-WADiS* competence test for scientific reasoning (Hartmann et al. 2015a; Mathesius et al. 2014; Stiller et al. 2016; Straube 2016).

In this chapter, we present evidence for the validity and reliability of our test score interpretations, discuss implications for the theoretical model and for the test instrument and its practical use, and provide an outlook for further use of the model and the test.

2 Theoretical Framework: Scientific Reasoning Competencies

Scientific reasoning (Giere et al. 2006; Klahr 2000) as a problem-solving process (Mayer 2007) is considered a key competence in basic science education for the natural sciences biology, chemistry, and physics (Rönnebeck et al. 2010, p. 178). As such it belongs to the indispensable core competencies for the 21st century (Trilling and Fadel 2009). Scientific reasoning competencies are cognitive dispositions

to gain empirical insights into scientific phenomena by successfully applying steps of an idealized problem-solving process to given scientific problems (Gut-Glanzmann and Mayer 2018; Mayer 2007). Within the three empirical natural sciences biology, chemistry and physics, central methods are scientific observation of phenomena, controlled experimentation by the variation and control of variables (Gut-Glanzmann and Mayer 2018; Mayer 2007; Wellnitz and Mayer 2013), as well as scientific modeling (Krüger et al. 2018; Upmeier zu Belzen and Krüger 2010).

In criterion-driven observations, competencies in predicting, describing, and systematically examining correlative relationships between structures and their functions under temporal changes are required (Wellnitz and Mayer 2013). Competencies in systematic experimentation reflect the ability to capture causal relationships with systematically varied and controlled conditions (Gut-Glanzmann and Mayer 2018; Mayer 2007). This requires the ability to handle independent and dependent variables as well as control variables (Roberts and Gott 2003). According to Mayer (2007), the scientific thinking processes underlying observations and experimentation as scientific investigations can be described as a domain-specific form of problem solving in four sub-competencies, which were operationalized in the Ko-WADiS competence test (Table 1): Formulating research questions, generating hypotheses, planning investigations, and analyzing and interpreting data (Gut-Glanzmann and Mayer 2018). These sub-competencies generally refer to experimentation, but can also be applied to observations, comparisons (Wellnitz and Mayer 2013), and the use of models (Upmeier zu Belzen and Krüger 2010). However, there are specific reasoning competencies needed in scientific modeling. Scientists, students, and teachers need to be able to reflect the role of models in the process of scientific modeling (Krüger et al. 2018). Thus, it has to be assessed whether and to which extent models are seen as tools to reconstruct central features of reality or to develop a methodological basis for the formulation of research questions (Gilbert and Justi 2016). According to Upmeier zu Belzen and Krüger (2010), using scientific models to reason about scientific phenomena is operationalized in sub-competences (Table 1) such as purpose of models, testing models, and changing models.

These seven sub-competencies of conducting scientific investigations and using scientific models (Table 1) are interrelated steps of a general scientific thinking process in the sense of the hypothetico-deductive approach (Popper 2004). Against the background of an ideal-typical view, a researchable scientific question is formulated on the basis of a real-life scientific problem (White 2017). Subsequently, a model is developed whose purpose is to generate inter-subjectively traceable and falsifiable hypotheses (Lawson et al. 2000). Testing these hypotheses means testing the model. For this, a suitable experimental arrangement must be planned

(Lawson et al. 2000). The results must be evaluated and interpreted and can either support or oppose the assumptions of the model. The latter option results in changing the model, which means that the process restarts. Being competent in the field of scientific reasoning is defined as a cognitive disposition that enables students to apply each of these seven steps to real-life scientific problems (Giere et al. 2006).

Table 1 Scientific reasoning competencies in the areas *conducting scientific investigations* and using *scientific models*

Competency	Scientific reasoning	
Dimension	conducting scientific investigations	using scientific models
Sub-competencies	formulating questions (18)	purpose of models (18)
	generating hypotheses (16)	testing models (18)
	planning investigations (22)	changing models (14)
	analysing data and drawing conclusions (17)	

Note. Number of developed items in brackets

The described sub-competencies become accessible to measurement by the cognitive-psychological construct of scientific reasoning (Fischer et al. 2014). In international research, they are also conceptualized in styles of scientific reasoning (Osborne 2018) or scientific paths of knowledge acquisition (Priemer et al. 2019). In science curricula (e.g., KMK 2014; NGSS 2013), the ability to carry out and reflect about scientific inquiry processes is mandatory.

3 The Ko-WADiS Competence Test

In the project, *Ko-WADiS*¹ (2011–2015), the paper-pencil *Ko-WADiS* competence test was developed (Hartmann et al. 2015a). The focus was on the clarification of the theoretical foundations, the development of the test instrument, the standardization of the items for the three subjects biology, chemistry, and physics (Mathesius et al. 2014), the investigation of the basic psychometric properties of the test, and the start of a longitudinal assessment in the target population (pre-service science teachers) as well as in various control groups (Hartmann et al. 2015a).

1 **Kompetenzmodellierung und -erfassung zum Wissenschaftsverständnis über naturwissenschaftliche Arbeits- und Denkweisen bei Studierenden**

The *Ko-WADiS* competence test is the result of a multi-step process (Hartmann et al. 2015b), in which students' responses in the open-ended format were checked for content validity in an expert discourse. The piloting of the multiple-choice items finally led to a set of 123 items (distribution see Table 1). To make the data collection and evaluation more economical, and to reduce the amount of missing values, 63 items with the best psychometric properties were selected from this item pool (three items per subject and sub-competence; Stiller et al. 2016). These items are used for longitudinal assessment since 2013. The paper-pencil-based single best answer items address scientific reasoning competencies as they are typical for pre-service teachers of biology, chemistry and physics (Hartmann et al. 2015a), but were also used by science students, students of psychology, and in-service biology teachers. Each test booklet contains 21 of the 63 items (balanced-incomplete block design; Gonzalez and Rutkowski 2010), and is assessed in 45 minutes. In each test booklet, the seven sub-competences and the three scientific disciplines are equally distributed. Because of the design, the test is evaluated using probabilistic methods.

4 Investigation of Validity

The validity of the *Ko-WADiS* test score interpretations was evaluated in a follow-up project, *ValidiDiS²* (2016–2019). Alongside a continuation of the longitudinal data collection to investigate competence development, we addressed different sources of validity evidence as described in the *Standards for Educational and Psychological Testing* (AERA et al. 2014; see also Kane 2013): evidence based on test content, evidence based on internal structure, evidence based on response processes, and evidence based on relations to other variables such as conceptually related constructs and criteria. Investigations on these sources of validity provide empirical evidence to support the assumption that the test results can be interpreted in terms of the underlying theoretical construct.

Finally, consequences of testing can serve as a potential source of validity evidence, but it was not used in our project as no consequences suitable for validity investigations are based on the test results yet.

2 **Kompetenzmodellierung und -erfassung: Validierung eines Modells zum wissenschaftlichen Denken im naturwissenschaftlichen Studium**

4.1 Evidence Based on Test Content

An investigation of content validity was addressed from the beginning of the *Ko-WADiS* project, thus starting prior to the development of the test instrument. To ensure a standardized item construction process which followed guidelines based on theoretical groundings (Mayer 2007; Upmeier zu Belzen and Krüger 2010), an item construction manual was developed. Answering options for the single-best answer items of the developed instrument were based on written answers of students to open-ended items (Hartmann et al. 2015b; Mathesius et al. 2014). Investigations of item features that systematically affect item difficulty revealed predictive potential for one formal item feature (length of response options), two features based on cognitive demands (processing data from tables, processing abstract concepts), and one feature based on solid knowledge (specialist terms). This was in accordance with the cognitive demands operationalized in the theoretical structure of the test. Thus, it is concluded that the findings support the validity of the interpretation of the test scores as measures of scientific reasoning competencies (Stiller et al. 2016).

Content validity was also examined by an expert rating. A sample of 21 academic teachers and researchers with a high level of theoretical knowledge and research experience in the field of scientific reasoning (Gruber 2010) were requested to classify the relationship between six selected items (two per scientific discipline) and the sub-competencies of the theoretical construct (Table 1). The experts correctly designated each item to the corresponding sub-competence, and evaluated how well the item represented the sub-competence on a five-point Likert scale with a value 1 for very poor and a value of 5 for very good theoretical resemblance of the item. The median rating was 4 out of 5 for five items, and 3 out of 5 for one item, with an interquartile range between 0.00 and 2.25. These results indicate an appropriate operationalization of the items, indicating that they represent the construct to a satisfactory degree (Hartmann et al. 2019a).

4.2 Evidence Based on Internal Structure

With respect to the internal structure, the empirical results from cross-sectional data support a one-dimensional structure of scientific reasoning, although the fit of a two-dimensional model that differentiates between aspects of scientific investigation and aspects of scientific modeling is not significantly worse than that of the one-dimensional model (Hartmann et al. 2015a). This is in accordance with the theoretical assumptions of the underlying construct, which assumes scientific

reasoning being generalizable across the subjects biology, chemistry and physics. Thus, the established unidimensionality indicates a broad theoretical construct. These results correspond with those of other empirical studies on scientific reasoning competencies (e.g., Mannel 2011; Neumann 2011; Wellnitz 2012).

4.3 Evidence Based on Response Processes

Response options (distractors and attractors) were generated out of students' written answers to open-ended questions, using the generated item stems as stimuli. This contributes to student-centered response options in the single best answer test and secures valid test score interpretations.

The investigation of response processes as a source of validity evidence was done by eyetracking collecting gaze data and verbal data while working on items of the Ko-WADiS competence test (12 items, $N = 16$; Mathesius et al. 2018). In addition to think-aloud protocols, the cued retrospective reporting method (attention maps, sequence charts; van Gog et al. 2005) was applied for the investigation of cognitive processes during eye tracking. Although pre-service biology teachers who chose the attractor do not differ in their eye movement patterns (fixations, dwell time) from those who chose the distractor, the verbal data describes the individual solution processes in a comprehensible way. The findings based on response processes are interpreted as evidence for the validity of the test scores interpretation as measures of scientific reasoning competencies (Mathesius et al. 2018).

4.4 Evidence Based on Relations to other Variables

Validity evidence based on relations to other variables was investigated utilizing several empirical methods (Table 2). An investigation of instructional sensitivity was accomplished by observing the development of test scores due to short-term learning progress in regular university seminars and lectures (1) and in intervention studies (2). Besides that, we investigated groups with either hypothesized mean differences (3) or mean equivalence (4). Finally, correlations with general abilities like intelligence and complex problem solving (5) and with conceptually related constructs, such as pre-service teachers' pedagogical content knowledge (6), were calculated.

1. Instructional sensitivity is supported by an investigation during regular academic training in courses of biology education: We used the long version of our instrument to test 59 pre-service biology teachers before and after a semester. Comparing the results, we found a moderate increase in the students' ability estimates ($t = 2.59$, $p_{\text{one-sided}} = .006$, $d = 0.34$). Instructional sensitivity was also investigated in an interventional study. A sample of 87 pre-service science teachers participated on a two-day intensive course to train scientific reasoning competencies. The instructional sensitivity was tested for a selection of nine items from our original instrument which were answered by the students before and after the intervention. The pre-post comparison of the sum score reveals a significant increase ($t = 2.30$, $p_{\text{one-sided}} = .012$, $d = 0.25$). In a control group ($N = 55$), no significant effect was found.

A sample of 125 pre-service biology teachers participated on a seminar to promote scientific reasoning competencies explicitly. The instructional sensitivity was tested with an item subset of 21 biology items before and after the intervention. The pre-post comparison of the sum score reveals a significant increase ($t = 4.72$, $p_{\text{two-sided}} < .001$, $d = 0.35$). In a control group ($N = 49$), no significant effect was found. These findings further support the interpretation of the test scores as measures of scientific reasoning competencies.

2. Known-group comparisons (Cronbach and Meehl 1955) are an economical tool to investigate aspects of criterion-based validity. We used it to test whether our test scores are sensitive to differences between undergraduate and postgraduate students, and to differences between pre-service science teachers who study one scientific discipline alongside a non-scientific discipline and pre-service science teachers who study two scientific disciplines. The comparisons were carried out as a latent regression analysis. The results support the hypotheses of group differences with significant regression effects of group affiliation on the latent ability measures ($B_{\text{academic phase}} = 0.283$, $p < .001$; $B_{\text{scientific disciplines}} = 0.116$, $p < .01$).

Additionally, known-group comparisons were carried out with 626 pre-service biology teachers (Mathesius et al. 2016). It was predicted that pre-service biology teachers who also study chemistry or physics perform better than pre-service biology teachers without a second science subject, and pre-service biology teachers in more advanced stages of academic education (study stages: 4th-7th semester and 8th-10th semester) perform better in the test than students in early stages (1st-3th semester). To test these hypotheses, multiple latent regression analysis was applied. The results show significant regression effects of group affiliation ($B_{\text{scientific disciplines}} = 0.774$, $p < .001$; $B_{\text{terms 4-7}} = 0.499$, $p < .001$; $B_{\text{terms 8-10}} = 1.279$, $p < .001$) on the latent ability measures. Both findings indicate

- that the test scores are sensitive to relevant criteria (Hartmann et al. 2015b; Mathesius et al. 2016).
3. In addition to the study of predicted mean differences, predicted mean equality was tested as well (Hartmann et al. 2019b). On the basis of initial grades, course and module descriptions, it was hypothesized that the levels of pre-service science teachers' and psychology students' scientific reasoning competencies do not show a meaningful difference. Therefore, the mean test scores in the two groups should be equivalent. To test the hypothesis, we compared the means of matched sub-samples with balanced covariate distributions, utilizing the two-on-one-sided-tests method (TOST; Schuirmann 1987) to test the equivalence of the students' abilities. The hypothesis of group equivalence is supported by the absence of a significant difference ($t = 0.03$; $p_{two-sided} = .98$) in combination with a very small effect size of $d = 0.00$ that is nominally below a pre-defined smallest effect size of interest ($d = 0.17$). However, the TOST procedure indicates that the confidence interval around the mean difference exceeds the equivalence bounds due to the relatively small sample size, rendering the results inconclusive ($t_{TOST} = 1.35$; $p_{one-sided} = .09$; Hartmann et al. 2019b).
 4. As a further indicator of validity, it was tested to what degree variance of the *Ko-WADiS* test scores can be attributed to more general skills such as intelligence or complex problem-solving abilities (Mathesius et al. 2019). The *Ko-WADiS* competence test scores and the scores of the reasoning scale of the Intelligence-Structure Test 2000 R (Liepmann et al. 2007) and complex problem solving (Genetics Lab test; Greiff and Fischer 2013; Sonnleitner et al. 2013) show positive significant correlations (I-S-T 2000 R: $r = .44$; $p_{two-sided} < .001$; Genetics Lab test: $r = .33-.40$; $p_{two-sided} < .001$). Furthermore, the regression analysis clarifies 24% of the variance by the considered variables. The findings support the assumption that they are distinct constructs with connecting facets. Therefore, part of the remaining variance might be interpreted as evidence for the test score interpretations as measures of scientific reasoning competencies (Mathesius et al. 2019).
 5. Finally, the convergent validity was investigated in a correlational study ($N = 65$ pre-service science teachers). We correlated sum scores from a short version of our instrument (15 items) with 12 scientific-reasoning items from an early testing version of the PCK-IBI (Großschedl et al. 2018). The scores from the two instruments correlate significantly ($r = .49$; $p_{one-sided} < .001$). Potential moderating effects of general cognitive ability were controlled by including the non-verbal subscale of the IST Screening (Liepmann et al. 2007) in the analysis, which left the correlation coefficient practically unchanged ($r_{partial} = .48$; $p_{one-sided} < .001$). The findings indicate that the correlation of the two tests is not explained by

intelligence, which provides further supporting evidence to the validity of our test score interpretations as measures of scientific reasoning competencies.

4.5 Summary

Overall, the outcomes of the validation studies are coherent and provide supporting evidence for the interpretation of the test scores as measures of scientific reasoning competencies. The majority of the investigated effects were small to medium, implying that the true effects are moderate. However, statistical power is limited due to mediocre reliabilities (Section 5), which certainly affect the power of the statistical procedures (Kanyongo et al. 2007).

Table 2 Overview of studies for sources of validity evidence

Source of validity evidence	Investigation of ...	Instrument and sample	Results	Reference
Test content	... process of item development	183 open ended items $N = 259$ 166 single best answer items $N = 578$	theory-based selection of 123 items based on item-parameter analysis	Hartmann et al. 2015b
	... item features affecting item difficulty	63 single best answer items $N = 907$; 9 experts	32 % of variance is explained by the item features investigated, and is in accordance with model assumptions and expert ratings from a standard setting	Stiller et al. 2016
	... expert ratings of selected test items	6 Likert-style items, 21 experts	selected items represent the theoretical construct to a satisfying degree	Hartmann, Krüger et al. 2019
Internal structure	... the dimensionality of the theoretical structure	141 items, $N = 3\ 010$	unidimensional structure in accordance with theoretical assumptions	Hartmann et al. 2015a; Hartmann et al. 2015b

Source of validity evidence	Investigation of ...	Instrument and sample	Results	Reference
Response processes	... gaze data and think aloud protocols	12 single best answer items $N = 16$	no correlation between selection of answer and eye movements, scientific reasoning is necessary to find the attractor	Mathesius et al. 2018
Relations to other variables	... the instructional sensitivity to progress in selected seminars and lectures	123 single best answer items, pre-post design, $N = 59$	moderate increase of scientific reasoning competencies ($d = 0.30$)	
		21 single best answer biology items, pre-post design, $N = 49$	increase of scientific reasoning competencies ($d = 0.54$)	Mathesius et al. in preparation
	... the instructional sensitivity in intervention studies	9 single best answer biology items, $N = 87$	increase of scientific reasoning competencies ($d = 0.25$)	Hartmann, Krüger et al. 2019
		21 single best answer biology items, $N = 125$	increase of scientific reasoning competencies ($d = 0.77$)	Mathesius et al. in preparation
	... groups with hypothesized mean differences	123 single best answer items, $N = 2247$	significant regression effects support hypothesized group differences	Hartmann et al. 2015b
		123 single best answer items, $N = 626$ pre-service biology teachers	positive effects of variables <i>number of natural sciences</i> (1 or 2) and <i>academic level</i> (Bachelor or Master) on test scores	Mathesius et al. 2016

Source of validity evidence	Investigation of ...	Instrument and sample	Results	Reference
Relations to other variables	... groups with hypothesized mean equivalence	63 single best answer items, $N = 184$	highly similar ability distributions but no significant equivalence effect	Hartmann, Ziegler et al. 2019
	... correlations with I-S-T-screening and the complex problem-solving micro world	123 single best answer items, I-S-T-screening, 12 GL-micro world problems, $N = 232$	24 % of variance are explained by the investigated variables	Mathesius et al. 2019
	... correlation with a conceptually related construct	15 single best answer biology items, 12 multiple-choice biology items, I-S-T-screening, $N = 65$	substantial correlation between the two instruments ($r = .49$) that remains stable if controlled for general cognitive ability ($r_{\text{partial}} = .48$)	Hartmann, Krüger et al. 2019

4.6 Competence Development

A general indicator for the sensitivity of the test scores on learning opportunities can be inferred from the longitudinal data we collected over the time span of pre-service science teachers' academic training at Freie Universität Berlin and Humboldt-Universität zu Berlin. The participants of the longitudinal *Ko-WADiS* study had to process the test at four times: at the beginning of their academic studies, in their fourth undergraduate semester, at the beginning of their postgraduate studies, and at the fourth postgraduate semester (which is usually the semester in which they graduate as a Master of Education). Data collection took place in regular academic seminars and lectures.

Utilizing a cohort-sequential longitudinal design, three complete cohorts with a total of 644 students were tested. Due to the balanced-incomplete block design, we utilized IRT models to estimate our students' competencies. WLE were used as measures of person ability. As students at the participating universities can choose freely in which semester they apply for certain courses, additional data collection

took place unsystematically at times different from the first and fourth semesters. Missing data in the longitudinal panel was imputed using the MICE procedure (multiple imputation with chained equations; van Buuren and Groothuis-Outshoorn 2011). The results show a moderate increase of competencies over time (Figure 1).

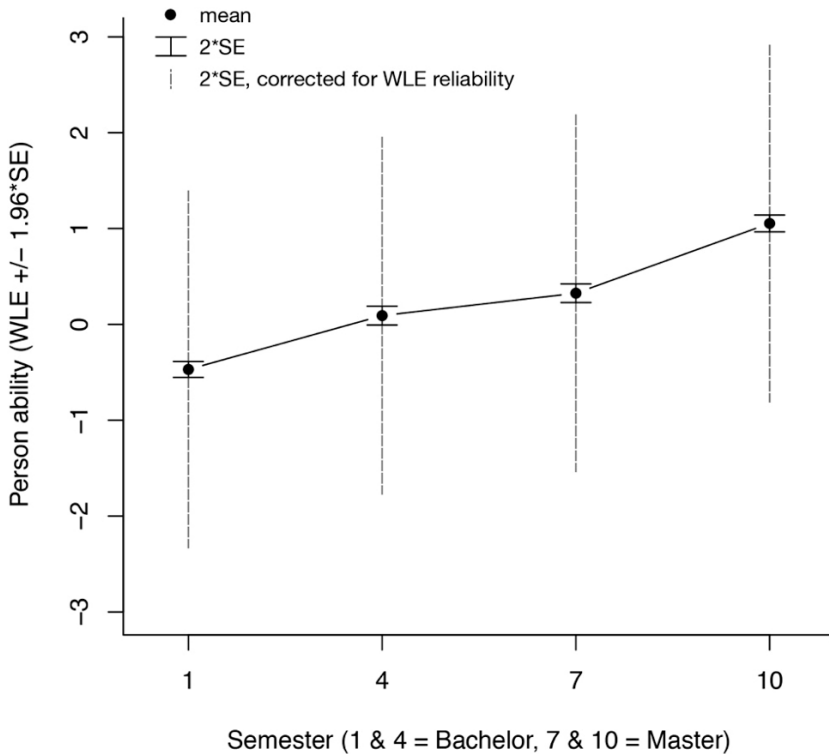


Figure 1 Development of 644 pre-service science teachers' scientific-reasoning skills during academic education. Means of weighted likelihood estimates (WLE), twofold standard errors, and twofold standard errors corrected for WLE reliability

Given that the competencies in question are part of the students' academic training, this increase is in accordance with our test score interpretation. However, the validity evidence that can be derived from this finding is limited, as other competencies increase during academic education as well.

5 Reliability

During the longitudinal *Ko-WADiS* assessment and the investigation of validity in the *ValiDiS* project, several person and item samples and subsamples were tested. Therefore, different measures of reliability were investigated. The reliability measures vary depending on the particular sample considered, but overall are satisfactory:

The rating scale used to investigate expert judgments had a *Cronbach's alpha* of .63 (six items, $N = 21$ experts). Dimensionality analyses were based on data from 3 010 pre-service science teachers and science students, utilizing a one-parametric logistic model with latent ability estimates. The according Expected-A-Posteriori/Plausible Value (EAP/PV) reliability is 0.47.

The test scores' instructional sensitivity to learning progress in regular lectures was investigated by comparing Weighted Likelihood Estimates (WLE) as measures of the 59 participants' individual abilities. The corresponding WLE reliability of the long version of the test (123 items) is 0.40, and the test-retest reliability is .60. Instructional sensitivity was further tested in an experimental intervention with 87 participants. The nine-item short version of the test has a *Cronbach's alpha* of .44 and a test-retest reliability of .48.

The test's sensitivity to known-groups differences was modelled as latent regression with latent estimates as measures for group means and variances. The corresponding EAP/PV reliability is 0.54 ($N = 2\ 247$; Hartmann et al. 2015b). In a subsample of pre-service biology teachers, the EAP/PV reliability of the test was 0.66 ($N = 626$; Mathesius et al. 2016).

Hypotheses of group equivalence were tested by comparing WLE distributions in two matched samples of 131 pre-service science teachers and 131 psychology students. The estimation was based on the optimized 63-item version of our instrument, with a WLE reliability of 0.59.

The short version of our instrument used to investigate the convergent validity (15 items) has a *Cronbach's alpha* of .61 ($N = 65$ pre-service biology teachers). The 21 item biology test booklet has a *Cronbach's alpha* of .60.

Investigating the empirical relationship between the *Ko-WADiS* test (120 items) with the I-S-T 2000 R and Genetics Lab test (Mathesius et al. 2019), the corresponding EAP/PV reliability was 0.55 ($N = 232$). Finally, the instrument used to investigate competence development (123 items, $N = 644$) has a WLE reliability of 0.50.

Overall, the reliabilities found in our studies are comparable to the values of other projects which reported reliabilities for scientific-reasoning tests between 0.23 and 0.66 (e.g., Mannel 2011; Neumann 2011; Wellnitz 2012).

6 Discussion and Implications

The *Ko-WADiS* competence test provides test takers, academic teachers and educational researchers with an instrument to measure competencies and their development reliably, validly and economically. Multiple evidences of validity support the interpretation of the test scores as a measure of scientific reasoning competencies. Longitudinal results show an increase in competence during academic education.

The intended use of the *Ko-WADiS* competence test is to provide an empirical basis to describe pre-service science teachers' scientific-reasoning competencies and the development of these competencies. The test was specifically designed for large-sample scenarios, such as monitoring studies. The use of the instrument for individual diagnoses is not intended and therefore has not been investigated. However, short versions of the test were used for the assessment of validity in relatively small samples (Hartmann et al. 2019).

A future perspective on test use might be the question how to further develop teaching and learning scientific reasoning competencies. We assume that, starting from our rather broad construct, learning opportunities focusing on specific aspects of scientific reasoning could be evaluated with short versions of the test. For example, the effects of a seminar about scientific modeling on students' modelling competencies could be investigated by utilizing a short version of our instrument that consists of test items from the modelling subscale. However, in such a scenario, the reliability and validity must be investigated again before inferences are drawn from the results.

In terms of dissemination such short-test might be used – eventually adapted – for the purpose of experimental intervention studies with a specific theoretical background and research questions. Such an approach would give additional insight into test characteristics, but at the same time would help to develop teaching and learning. Digitalization of the test – eventually in an adaptive way – might help dissemination (Brügge- man and Nordmeier 2018). The dissemination of the test use or the use of short-tests means to evaluate possible transfer of scientific reasoning competencies.

The effectiveness of seminars fostering scientific reasoning competencies can also be investigated in interventions using the *Ko-WADiS* test (Mathesius et al. in preparation). The *Ko-WADiS* test has not been developed for individual diagnosis. Nevertheless, a computer-aided adaptive test with three test blocks of five items each is in preparation to enable individual diagnose. The goal is to make the measurement of scientific reasoning competencies more economical while maintaining the same reliability and validity, and to enable individual diagnostics (Brügge- man and Nordmeier 2018).

Limitations arise from the rather low reliability of the test scores. Though not unusual for tests of this kind, the mediocre consistency measures indicate a notable amount of standard error, which significantly affects the outcome of inferential analyses. Paired with the relatively small effect sizes we found in almost all scenarios, the potential of using the test in small samples is limited.

The discussion of small effect sizes brings up two different strands of argumentation. Either, the assumption of unidimensionality of the scales goes along with low reliabilities that are explained by construct-irrelevant variance. Following Cronbach (1951), the reliability should be calculated separately for the items relating to different sub-dimensions. In this case, we would face the problem of construct-underrepresentation as we wouldn't have enough items in each test booklet. Therefore, a possible interpretation for issues of reliability of a research instrument could be that students' responses are highly situated and contextualized (Leach et al. 2000). Adams and Wieman (2011) pick up on this point by arguing that a high correlation between tasks means that the tasks are repetitive. The observation that in the preceding validation analyses students understood the items as intended and gave reasonable explanations for their responses (Mathesius et al. 2018) could indicate that the low reliability is a consequence of the students' diverse understanding across the sub-dimensions. However, the dimensionality analyses as well as correlations between items and between sub-competences do not foster this interpretation (Hartmann et al. 2015a).

With regard to international analyses, the 21 biology items of the *Ko-WADiS* instrument were translated into English, Spanish and Greek (TRAPD: Translation, Review, Adjudication, Pretest, Documentation, Harkness et al. 2010) and assessed in Australia (Krell et al. 2018), Chile (Krell et al., in preparation), Canada³ and Cyprus. The translation into French is currently taking place. As far as investigated, only a few, already revised items in other languages led to moderate DIFs (Krell et al. 2018).

References

- Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33(9), 1289–1312.
- American Educational Research Association, American Psychological Association & National Council on Measurement in Education [AERA, APA & NCME] (2014). *Standards*

3 Ethics committees in Germany and Canada have approved the questionnaire for use.

- for educational and psychological testing. Washington, DC: American Educational Research Association.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Brüggemann, V., & Nordmeier, V. (2018). Naturwissenschaftliches Denken im Lehramtsstudium- Computeradaptive Leistungsmessung. In C. Maurer (Ed.), *Qualitätvoller Chemie- und Physikunterricht – normative und empirische Dimensionen*. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung in Regensburg 2017 (pp. 915–918). Regensburg: Universität Regensburg.
- Van Buuren, S., & Groothuis-Outshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, 2(3), 28–45.
- Giere, R. N., Bickle, J., & Mauldin, R. F. (2006). *Understanding scientific reasoning*. Independence, KY: Wadsworth/Cengage Learning.
- Gilbert, J. K., & Justi, R. (2016). *Modelling-based teaching in science education* (Vol. 9). Switzerland: Springer.
- Van Gog, T., Paas, F., van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *Journal of experimental psychology*, 11(4), 237–244.
- Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph*, 3, 125-156.
- Greiff, S., & Fischer, A. (2013). Der Nutzen einer komplexen Problemlösekompetenz. *Zeitschrift für Pädagogische Psychologie*, 27(1-2), 27–39.
- Großschedl, J., Welter, V., & Harms, U. (2018). A new instrument for measuring pre-service biology teachers' pedagogical content knowledge: The PCKIBI. *Journal of Research in Science Teaching*. Advance online publication. Doi:10.1002/tea.21482
- Gruber, H. (2010). Expertise. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (pp. 183–189). Weinheim: Beltz.
- Gut-Glanzmann, C., & Mayer, J. (2018). Experimentelle Kompetenz. In D. Krüger, I. Parchmann & H. Schecker (Eds.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (pp. 121–140). Berlin: Springer.
- Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.Ph., Pennell, B.-E., & Smith T.W. (Eds.). (2010). *Survey methods in multinational, multiregional, and multicultural contexts*. New Jersey: John Wiley & Sons.
- Hartmann, S., Mathesius, S., Stiller, J., Straube, P., Krüger, D., & Upmeyer zu Belzen, A. (2015a). Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung als Teil des Professionswissens zukünftiger Lehrkräfte. In B. Koch-Priewe, A. Köker, J. Siefried &

- E. Wuttke (Eds.), *Kompetenzerwerb an Hochschulen: Modellierung und Messung* (pp. 39–58). Kempten: Klinkhardt.
- Hartmann, S., Upmeier zu Belzen, A., Krüger, D., & Pant, H. A. (2015b). Scientific reasoning in higher education: Constructing and evaluating the criterion-related validity of an assessment of preservice science teachers' competencies. *Zeitschrift für Psychologie*, 223, 47–53.
- Hartmann, S., Krüger, D., & Upmeier zu Belzen, A. (2019a). Investigating the validity and reliability of a scientific reasoning test for pre-service teachers. Vortrag auf der ESERA September 2019, Bologna.
- Hartmann, S., Ziegler, M., Krüger, D., & Upmeier zu Belzen, A. (2019b). "Equivalent-groups validation"? Practical application and critical examination of a known-groups approach to investigate the criterion-related validity of test score interpretations.
- Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42, 448–457.
- Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, 6, 81–90. Doi:10.22237/jmasm/1177992480
- Klahr, D. (2000). Exploring science. *The cognition and development of discovery processes*. Cambridge, MA: MIT Press.
- KMK (Ed.). (2014). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung*. Berlin, Germany: author. Retrieved from http://www.akkreditierungsrat.de/fileadmin/Seiteninhalte/KMK/Vorgaben/KMK_Lehrerbildung_inhaltliche_Anforderungen_aktuell.pdf
- Koepfen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie*, 216, 61–73.
- Krell, M., Redman, C., Mathesius, S., Krüger, D., & van Driel, J. (2018). Assessing pre-service science teachers' scientific reasoning competencies. *Research in Science Education*, 1–25.
- Krell, M., Mathesius, S., van Driel, J., Vergara, C., & Krüger, D. (in preparation). Assessing scientific reasoning competencies of pre-service science teachers: Applying the TRAPD approach to translate a German multiple choice questionnaire into English and Spanish. *International Journal of Science Education*.
- Krüger, D., Kauertz, A., & Upmeier zu Belzen, A. (2018). Modelle und das Modellieren in den Naturwissenschaften. In D. Krüger, I. Parchmann & H. Schecker (Eds.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (pp. 141–157). Berlin: Springer.
- Lawson, A.E., Clark, B., Cramer- Meldrum, E., Falconer, K.A., Sequist, J. M., & Kwon, Y.-J. (2000). Development of scientific reasoning in college biology: Do two levels of general Hypothesis-testing skills exist? *Journal of Research in Science Teaching*, 37, 81–101.
- Leach, J., Millar, R., Ryder, J., & Séré, M.-G. (2000). Epistemological understanding in science learning: the consistency of representations across contexts. *Learning and Instruction*, 10(6), 497–527.
- Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Göttingen: Hogrefe.
- Mannel, S. (2011). *Assessing scientific inquiry. Development and evaluation of a test for the low-performing stage*. Berlin: Logos.

- Mathesius, S., Upmeyer zu Belzen, A., & Krüger, D. (2014). Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung. *Erkenntnisweg Biologiedidaktik*, 13, 73–88.
- Mathesius, S., Hartmann, S., Upmeyer zu Belzen, A., & Krüger, D. (2016). Scientific reasoning as an aspect of pre-service biology teacher education: Assessing competencies using a paper-pencil test. In T. Tal & A. Yarden (Eds.), *The future of biology education research* (pp. 93–110). Haifa, Israel: The Technion, Israel Institute of Technology/The Weizmann Institute of Science.
- Mathesius, S., Upmeyer zu Belzen, A., & Krüger, D. (2018). Eyetracking als Methode zur Untersuchung von Lösungsprozessen bei Multiple-Choice-Aufgaben zum wissenschaftlichen Denken. In M. Hammann & M. Lindner (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik: Band 8* (pp. 225–244). Innsbruck: Studienverlag.
- Mathesius, S., Krell, M., Upmeyer zu Belzen, A., & Krüger, D. (2019). Überprüfung eines Tests zum wissenschaftlichen Denken unter Berücksichtigung des Validitätskriteriums relations-to-other-variables. *Zeitschrift für Pädagogik*, 65(4), 492–510.
- Mathesius, S., Bruckermann, T., Schlüter, K., & Krüger, D. (in preparation). Assessing pre-service science teachers' scientific reasoning competencies: Using known-groups as source of validity evidence for a scientific reasoning test.
- Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Eds.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (pp. 177–186). Berlin Heidelberg: Springer.
- Neumann, I. (2011). *Beyond physics content knowledge. Modeling competence regarding nature of science inquiry and nature of scientific knowledge*. Berlin: Logos.
- NGSS Lead States (Ed.). (2013). *Next generation science standards: for states, by states*. Washington, DC: The National Academies Press.
- Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity*, 10, 265–279.
- Osborne, J. (2018). Styles of scientific reasoning: What can we learn from looking at the product. Not the process, of scientific reasoning? In F. Fischer, C. A. Chinn, K. Engelmann & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 162–186). New York: Taylor & Francis.
- Popper, K. R. (2004). *Unended quest: An intellectual autobiography*. London: Routledge.
- Priemer, B., Eilerts, K., Filler, A., Pinkwart, N., Rösken-Winter, B., Tiemann, R., & Upmeyer zu Belzen, A. (2019). A framework to foster scientific problem-solving in STEM and computing education. *Research in Science & Technological Education*, 3(2), 1–26.
- Roberts, R., & Gott, R. (2003). Assessment of biology investigations. *Journal of Biological Education*, 37(3), 114–121.
- Rönnebeck, S., Schöps, K., Prenzel, M., Mildner, D., & Hochweber, J. (2010). Naturwissenschaftliche Kompetenz von PISA 2006 bis PISA 2009. In: E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Eds.), *PISA 2009 Bilanz nach einem Jahrzehnt* (pp. 177–198). Münster: Waxmann.
- Schuurmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15, 657–680.
- Sonnleitner, P., Keller, U., Romain, M., & Brunner, M. (2013). Students' Complex Problem-solving Abilities. *Intelligence*, 41(5), 289–305.

- Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment and Evaluation in Higher Education*, 41(5), 721–732.
- Straube, P. (2016). *Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-)Studierenden im Fach Physik*. Berlin: Logos-Verlag.
- Trilling, B., & Fadel, C. (2009). *Twenty-first century skills. Learning for life in our times*. San Francisco: Jossey-Bass.
- Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 16, 41–57.
- Wellnitz, N. (2012). *Kompetenzstruktur und -niveaus von Methoden der naturwissenschaftlichen Erkenntnisgewinnung*. Berlin: Logos.
- Wellnitz, N., & Mayer, J. (2013). Erkenntnismethoden in der Biologie – Entwicklung und Evaluation eines Kompetenzmodells. *Zeitschrift für Didaktik der Naturwissenschaften*, 19, 315–345.
- White, P. (2017). *Developing research questions*. London: Palgrave Macmillan.



3.6

Performance Assessment of Generic and Domain-Specific Skills in Higher Education Economics

Nagel, M.-T., Zlatkin-Troitschanskaia, O., Schmidt, S., and Beck, K.

Abstract

Following criticisms by employers about academic graduates' lack of 21st century skills, students need to develop skills such as professional knowledge, critical thinking and problem solving. Accordingly, there is a demand for suitable assessments of these skills. One approach is to develop a performance assessment using tasks adapted from real-world decision-making and judgment situations that students and graduates have to face in academic and professional domains. Such tasks employ real-life scenarios and require generic and domain-specific skills in different facets to handle a given problem adequately. In this paper, we present a newly developed performance assessment that aims to measure such skills among higher education economics students and graduates of economics and we report results from two validation studies.

Keywords

Generic skills, domain-specific knowledge, critical thinking, graduates of economics, performance assessment, validation

Funding Details

This work was supported by the German Federal Ministry of Education and Research with the funding number 01PK15001A.

1 Introduction

In Germany, we are still lacking performance assessments that meet the methodological requirements for measuring university students' generic higher-order cognitive skills and that further meet the demands of the curriculum-instruction-assessment triad (Pellegrino et al. 2001) in higher education (Zlatkin-Troitschanskaia et al. 2018a). Initial attempts to adapt and validate existing performance tasks on critical thinking and problem solving from the U.S. for German contexts revealed significant limitations (for the adaptation and validation of CLA+ tasks in Germany, see Zlatkin-Troitschanskaia et al. 2018b). In particular, transferring the constructs underlying the tasks as well as developing the according scoring rubrics, which are necessary to rate the students' performance in critical thinking, have been challenging. It turned out that, for instance, cultural differences presented us with vast problems of interpretation and comparison.

Therefore, we developed an innovative performance assessment learning (PAL) task (Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b) to measure these skills among higher education economics students and graduates in Germany. The PAL task consists of a realistic short-frame scenario, where test takers are confronted with the succinct description of a situation and a resulting problem. The scenario is complemented by a document library of additional background information that varies in relevance, reliability, credibility and validity. Test takers are asked to react to the presented problem by using this information and writing a well-founded recommendation for action (Section 3).

In this paper, we present first results of an assessment of university students of economics using the PAL task. To control for their domain-specific knowledge, we used the WiWiKom test (Zlatkin-Troitschanskaia et al. 2019a; see also Schlax et al. in this volume). Student general cognitive ability was assessed by the intelligence test IST-2000 R (Liepmann et al. 2007). In addition, the test takers' final school-leaving grades as well as their grades in attended study modules in higher education economics were assessed. This study design allows for measuring and investigating the relationships between different facets of domain-specific and generic skills.

2 Conceptual Background and Research Hypotheses

Critical thinking is receiving attention as a 21st century skill that is internationally considered indispensable for students of all disciplines who want to be successful not only in the national, but also in the global context after graduation (Allgood and Bayer 2016; Allgood et al. 2015). Due to global economic change and the increasing internationalization of markets, critical thinking skills are considered an ever more important requirement (e.g., Willingham 2007; McGoldrick and Garnett 2013) – particularly for business and economics professions but also for the public.

Critical thinking is described in literature as a complex, multi-dimensional construct that often includes the elements problem solving, communication ability, media literacy, and other 21st century skills. Thus, without further clarification, the concept seems to be quite vague (e.g., Lai and Viering 2012). For the context of this study, we develop a working definition below.

While the importance of such higher-order cognitive skills is commonly accepted, they are not yet systematically taught in higher education (e.g., Browne et al. 1995; Arum and Roksa 2010). A current nationwide analysis of 32 German higher education degree programs and module descriptions in economics showed that, although critical thinking is considered an objective of teaching and learning outcomes in the respective curricula, it is generally not implemented explicitly or actively on the instructional side (Zlatkin-Troitschanskaia et al. 2018b). According to the curriculum-instruction-assessment triad (Pellegrino et al. 2001), this is predominantly due to a lack of learning tools (i.e. appropriate methods to teach critical thinking) and corresponding testing instruments for conveying, assessing and fostering such skills in economics (Hoyt and McGoldrick 2012).

In our study, we follow a holistic understanding, assuming critical thinking not to be the sum of other individual skills, but an integration of various sub-capabilities (Shavelson et al. 2018a; Zlatkin-Troitschanskaia et al. 2019b). Based on the current state of research on critical thinking (e.g., Halpern 2014; Liu et al. 2014; Paul and Elder 2006; Facione 2000), we developed a systematical synthesis in which we differentiate between the following key dimensions and their central subdimensions to operationalize *critical thinking* and thereby form a basis for the assessment of higher education students (Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b):

1. Evaluating and using information and sources in terms of relevance to the argument, reliability, validity, credibility of sources;

2. Recognizing, evaluating and using arguments and their components (such as claims, support, beliefs, assumptions or proven facts) with regard to evidentiality, objectivity, validity, consistency;
3. Developing sound and valid arguments based on information provided in the task, integrating additional information into coherent arguments, and structuring arguments consistently. This includes avoiding logical inconsistencies, identifying omissions and weaknesses, and evading decision-making errors or biases (e.g., due to “fast thinking” in contexts that call for “slow thinking”; Kahneman 2011; West et al. 2008);
4. Recognizing main and ancillary effects and evaluating consequences of decision-making and associated actions;
5. Appropriately communicating the most suitable course of action based on the given task prompt, i.e., making an evaluative judgment, explaining a decision, recommending a course of action by suggesting a solution to a problem.

Taking into account these facets and the various existing descriptions of critical thinking, the following working definition was developed for the study:

Students with advanced critical thinking skills question existing assumptions and opinions, and recognize and evaluate the relevance and reliability of provided information. Based on this judgement, logical or causal interrelationships are identified, consequences are considered and conclusions are drawn. Consequently, own arguments and opinions are formed, which in turn are reflected upon and corrected if necessary. Finally, the arguments and conclusions are communicated (written or verbally) in an understandable and convincing way.

The PAL task is designed to measure the critical thinking skills of higher education students or graduates according to the aforementioned working definition and can be applied to different domains including economics, as the context of the task relates to socio-economic topics. The PAL task builds on prior research on generic higher-order cognitive skills and, in particular, on the concept of critical thinking, which – in short – is described as the ability to analyze problems, evaluate claims and information, draw inferences, and weigh decisions with regard to their consequences. The task represents the next generation of performance assessments due to its innovative approach to performance assessment (Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b).

This newly developed PAL task was comprehensively validated in two subsequent studies in Germany (Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b) in accordance with the internationally established Standards for Educational and Psychological Testing by AERA et al. (2014). In this paper, we focus on the valida-

tion criterion ‘relation to other variables’, i.e., convergent and discriminant validity, to examine the relationships between different facets of domain-specific and generic skills: critical thinking, domain-specific knowledge and general cognitive ability. Using the method of comparing known groups (Hattie and Cooksey 1984) we also investigate the specificity and sensitivity of the newly developed PAL task for the *domain of economics*.

The research literature remains inconclusive in terms of the extent to which critical thinking is a domain-specific or domain-independent higher-order ability. It is often hypothesized that critical thinking itself is a generic skill, which can, however, only be conveyed and learned in concrete domain-specific contexts (e.g., Fives and Dinsmore 2017). The ability to perceive and process a domain-specific problem such as the scenario presented in the PAL task as well as perform the according solution requires a substantial level of expertise in this domain, since a profound understanding of the subject area is necessary for the handling of such a complex task (Alexander et al. 2016; Pellegrino and Hilton 2012). Accordingly, *graduate students (master), who should have greater domain-specific expertise, can be expected to perform at a higher level in terms of their critical thinking abilities in the domain of economics than undergraduate students (bachelor) (Hypothesis 1)*.

The PAL task focuses on critical thinking in a socio-economic context requiring not only domain-specific knowledge and an understanding of basic economic principles but also domain-independent higher-order skills, in terms of discriminant validity (e.g., Messick 1989). Therefore, we expect *a positive but relatively weak correlation between the critical thinking performance of students as measured by the PAL task and their performance in the domain-specific economic knowledge WiWiKom test (Hypothesis 2)*.

As PAL also measures other higher-order cognitive abilities besides domain-specific knowledge, which are needed to complete this holistic task, it can be assumed *that a good performance in the PAL task is also slightly positively correlated with general cognitive abilities (Hypothesis 3)* (which were assessed through intelligence sub-tests from the task groups “Choosing figures” and “Matrices” from the German intelligence test IST-2000 R as well as final school-leaving grades).

To further examine the domain specificity of PAL, we also applied the method of comparing known groups by comparing the results of students with a major in economics with those of students without a major in economics. *We expect the students in economics to do at least slightly better in the PAL task than students enrolled in other subjects (Hypothesis 4)*.

In addition, the influence of prior domain-specific knowledge, which might be due to previous education with an economic focus (e.g., completed commercial vocational training) was also controlled for.

3 Methods

3.1 Sample

For Germany, two samples have been surveyed within two subsequent validation studies that took place in the winter semester of 2017/2018 and the summer semester of 2018. Overall, 55 students from a German university participated – 25 undergraduates (bachelor's degree) and 30 graduates (master's degree). For their participation in the voluntary validation studies students could choose between an incentive of €20 or credits for a study module.

We contacted all 44 master's students enrolled in economics education at this university inviting them to participate in the study. Thus, we have a 68 % participation rate of all economics education master's students. Taking into account the sociodemographic characteristics, this sub-sample can be considered representative for this study domain (Tables 1a and 1b).

The procedure for recruiting bachelor's students was the same as for master's students, although in this case we managed to encourage slightly less than half of all students enrolled at the university in this course of study to participate. Based on the descriptive characteristics of the 25 participating bachelor's students and a descriptive comparison of characteristics with the results of the Germany-wide representative WiWiKom study with bachelor's students, this sub-sample can also be considered representative (Zlatkin-Troitschanskaia et al. 2019a; see also Schlax et al. in this volume). Since about half of the test takers were in the last year of their bachelor's studies and most of the graduate students were in the last year of a master's degree, this study not only provides preliminary data on advanced undergraduates' critical thinking skills but also indicates the level of critical thinking in graduate students towards the end of their studies.

Tables 1a and 1b show the descriptive statistics of the sample.

Table 1a Sample description

Attributes	<i>N</i> = 55
Gender	
Female	38 (69.1 %)
Male	17 (30.0 %)
Degree	
Bachelor	25 (45.3 %)
Master	30 (54.5 %)
Degree course	
Economic studies	46 (83.7 %)
Other	9 (16.3 %)
Completed vocational training	
Yes	23 (41.8 %)
No	30 (54.5 %)
Not specified	2 (3.6 %)
Completed internship	
Yes	42 (76.4 %)
Not specified	9 (16.4 %)

Table 1b Sample description (continued)

Attributes	<i>N</i> = 55		
	<i>N</i>	Mean	Std. Dev.
Age	54	24	3.44
Semester (Bachelor)	25	4.36	1.47
Semester (Master)	30	2.03	1.22
University entry qualification grade*	54	2.21	0.60

Note. *Grades vary from 1.0 (best) to 5.0 (worst).

3.2 Test Instruments and Administration

The PAL task, called “Wind Turbine”, has been constructed from authentic alternative energy source cases with meaningful consequences for a myriad of actors depending on the decisions and actions taken (Shavelson et al. 2019). More precisely, it focuses on the decision of a small-town council regarding whether or not to acquire and set up wind turbines on communal land near the town. Test-takers are presented with a document library and are asked to evaluate available information (e.g., a newspaper article, web documents, wind turbine schematics, stake-

holder interests as well as selected technical, economic, territorial law, settlement and wildlife data) that varies in terms of reliability, validity, and risk of bias or judgmental error. The task prompt requires participants to write an argumentative statement and recommend a course of action whether or not to set up the wind turbines and which further measures to take (for examples, see Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b). Test-takers are asked to use only the information provided and told that there is no right or wrong answer but that answers can vary in terms of their justifiability. They are also informed of the main scoring criteria. In addition to judging the trustworthiness and relevance of the different library documents, test-takers need to develop arguments for or against wind turbine construction. This process requires them to assess the value of each document while taking into account possible bias or motives for hidden agendas, such as personal profit and consequences for the community or individual residents. Task difficulty is fine-tuned by the nature of the information presented (reliability, validity, bias/error), the number of information sources, and the various points to consider by the test taker (e.g., stakeholder interests, consequences of the decision).

A rating scheme for scoring the written responses for the wind turbine PAL task has been developed based on the previously described definition of critical thinking (for details, see Zlatkin-Troitschanskaia et al. 2019b). It divides the students' response texts into four main dimensions with 4–9 performance criteria each (23 categories in total), whereby the facets 1–3 were grouped into two dimensions for reasons of formulating adequate behavior anchors. Then, by assessing individual PAL responses, 6 scoring anchors were formulated; on this basis, the participants' task performance can be assessed on a scale of 1–6 (for rubric 1 as an example, see Shavelson et al. 2019).

After rater training and a random designation of the individual PAL responses to two of a total of four raters, the responses were assessed using the developed rating scheme resulting in every task response of each participant being independently assessed by two raters.¹ To allow for a comparison of average performance results between the dimensions in spite of their different number of sub-categories and thus of attainable assessment points, the score of every dimension was divided by the number of its performance criteria. As a result, the calculated mean scores of every dimension can vary on a scale from 1 (requirements not fulfilled) to 6 (requirements fulfilled).

To assess knowledge and understanding of economics, we employed the *WiWiKom* test, which was validated in the representative, Germany-wide

1 There are sufficient inter-rater reliabilities between .7 and .85 (for further details, see Shavelson et al. 2019).

WiWiKom study with over 9,000 economics students according to the Standards for Educational and Psychological Testing (Zlatkin-Troitschanskaia et al. in 2019a; see also Schlax et al. in this volume). The WiWiKom test combines 15 items from the adapted and validated German version of the standardized *Test of Economic Literacy, Fourth Edition (TEL IV)* (Walstad, Rebeck, and Butters 2013) and 10 items from the adapted and validated German version of the *Test of Understanding in College Economics, Fourth Edition (TUCE IV)* (Walstad et al. 2007; for validation and adaptation in Germany, see Zlatkin-Troitschanskaia et al. 2014). The items of TEL IV operationalize basic principles of economics (such as the supply-demand model), complemented by five TUCE IV items from the microeconomics part and five TUCE IV items from the macroeconomics part. Every question offered 4 possible answers in multiple-choice format, with only one correct option.

The figural-spatial intelligence as indicator of general (fluid) intelligence of the students was measured using the two task groups “Choosing figures” and “Matrices” from the German intelligence test IST-2000 R (Liepmann et al. 2007). In total, this test includes 12 task groups, of which the two aforementioned were considered most suitable as good indicators of general intelligence (Liepmann et al. 2007). Each scale consists of 20 tasks. The participants have 7 minutes to complete the tasks in “Choosing figures” and 10 minutes for “Matrices”. In the former task group, students have to work out which of five given figures can be created by piecing together ten fragments. In the latter, students are given figure matrices built according to a certain rule and have to decide which of the five figure options would complete the matrix according to the rule.

The PAL task as well as the WiWiKom test were computer- and online-based. In addition to the students’ written answers, some data on the response processes (required time, information used for problem solving) were also collected and included in the analyses. The intelligence test was administered on-site via paper-pencil questionnaires under controlled conditions to ensure that the task was carried out properly.

Further socio-demographic information expected to affect test performance was collected as well. Prior studies have shown that task readability impacts the test performance of migrant students (e.g., Happ et al. 2019) and was therefore systematically controlled for in our study. Furthermore, several studies have demonstrated the suitability of the higher education entrance qualification as a reliable indicator for generic cognitive skills (e.g., Kobrin et al. 2008). Other indicators of relevant expertise in the context of solving the PAL task, such as completed commercial or vocational training, were surveyed as well as they might influence task performance and should therefore be considered.

The total test time for the PAL task is 60 minutes plus about 60 minutes for the other tests and questionnaires. The limited time puts the participants under pressure, requiring them to limit their focus by choosing the most relevant documents from the library and narrowing down their arguments. This leads to the active decision to discard certain information without considering it as otherwise the task would not be completed in time. The test-takers completed the task on computers provided by the project coordinators. In addition to the students' written answers, further data was collected on the response processes (e.g., behavior while working on the tests)², which was part of different analyses.

4 Results

To test *Hypothesis 1* (H1), expecting that graduate students (master) perform at a higher level than undergraduate students (bachelor), we conducted various analyses based on the written and scored responses to the PAL task. The lengths of students' responses vary significantly: the longest response comprises 1365 words and the shortest 68 words, with a mean word count of 495 ($SD = 218.7$). A significant difference can also be seen in the length of the answers between the bachelor's and master's students, with the master's students writing longer answers on average: The t-test yields $p = 0.0185$, mean bachelor = 421.08 ($SD = 200.50$), and mean master = 561 ($SD = 216.31$).

The analysis of the test performance shows that the length of the responses (number of words in the written statement) significantly correlates with the assessed test result measurements of the respective texts (Pearson correlation $r = 0.6$, $p = .00$). The master's students accordingly perform better on average than the bachelor's students (Table 3).

On average, the participants scored 83 out of a maximum of 138 points, with a minimum of 32 points and a maximum of 117 points. The score distribution has a skewness of -1.5 and kurtosis of 7.2, indicating test scores slightly skewed to the left. Although none of the participants earned a perfect overall score, some earned a full six points on some of the individual dimensions. Bachelor's students' average score was 81 points, with a minimum of 32 points and a maximum of 107 points, whereas the master's students achieved 85 points on average with a minimum of

2 For instance, the observation of the test-takers captured a wide range of approaches, with one extreme being participants investigating all the provided links to external information, whilst students on the other extreme only took into account the information provided on the task sheet itself.

56.5 points and a maximum of 117 points. This shows a higher level of critical thinking ability in graduate students and thus, at the descriptive level, the results are in line with our assumption (H1). Observing, however, the quartile level with respect to both groups, most graduate students place within the third quartile and only few in the fourth quartile of attainable points. Table 2 shows the distribution of students' PAL scores subdivided in quartiles. In spite of the obvious differences between performance scores of both groups, a t-test does not yield significant results, with $p = 0.15$, mean bachelor = 81.12 ($SD = 16.84$), and mean master = 87.6 ($SD = 16.04$).

Table 2 Distribution of students' PAL scores in quartiles

Quartile	Bachelor		Master	
	<i>N</i>	%	<i>N</i>	%
1 (0 – 34.5 points)	1	4 %	-	-
2 (35 – 69 points)	3	12 %	5	17.24 %
3 (69.5 – 103.5 points)	20	80 %	19	76 %
4 (104 – 138 points)	1	4 %	5	17.24 %

Further analyses of the four subdimensions reveal in which subskills students, on average, have better or worse performance. Table 3 shows a comparison of the average PAL subscale scores of bachelor's vs. master's students. In all subscales, master's students perform better than bachelor's students. However, both groups score lowest in subscale 3, "Recognizing and evaluating consequences of decision-making and actions", with average scores below 3 points, which constitutes less than half of the points that could be reached and again underlines the difference to the other sub-scales. According to the working definition, this is a vital facet of critical thinking, as in comparison to the other facets it requires students to apply their highest cognitive and meta-cognitive abilities.

Bachelor's students achieved their best results in subscale 1, "Recognizing and evaluating the relevance, reliability, and validity of given information", in which both groups scored approximately 4 points; therefore, compared to the other categories, all of the students' abilities rank high for this subscale. The same holds true for most students with regard to subscale 4, "Writing effectiveness and mechanics", in which the master's students achieved their best results (bachelor: 3.8 points, master: 4.2 points). The main difference between the two samples might be traced back to training in the context of bachelor's theses. The bachelor's thesis

is the first opportunity in the course of study for which a longer, more systematic scientific text must be written, which could significantly enhance students' writing skills. For subscale 2, "Evaluating and making a decision", in comparison to the other subscales, both groups' mean results are rather average, with only a small difference between bachelor's (3.5) and master's (3.8) students.

Table 3 Group performances for the PAL subscales

Subscale	Group Average Performance in PAL	
	Bachelor	Master
1: Recognizing and evaluating the relevance, reliability, and validity of given information	4.03 (SD = .70)	4.15 (SD = .62)
2: Evaluating and making a decision	3.47 (SD = .80)	3.81 (SD = .75)
3: Recognizing and evaluating consequences of decision-making and actions	2.69 (SD = .84)	2.87 (SD = .79)
4: Writing effectiveness and mechanics	3.84 (SD = .92)	4.20 (SD = .95)

Overall, the findings from the four subdimensions indicate a significant difference between undergraduate and graduate students only for „Writing effectiveness and mechanics“. With regard to the three abovementioned subdimensions 1, 2, and 3, students only show a high level of underlying abilities, indicated by high performance scores, in one subdimension, "Recognizing and evaluating the relevance, reliability, and validity of given information". In the two other subdimensions, performance levels are remarkably low in comparison, particularly for graduate students. Based on these findings, H1 cannot be rejected, as – even if only marginally in some cases – overall the graduate students performed better than the undergraduates.

As described in *Hypothesis 2* (H2), due to the economics-related context of the PAL task, it can be inferred that successfully solving the PAL task correlates with a high level of domain-specific knowledge and economic expertise. The 25 items of the WiWiKom test were evaluated as either incorrect (0 points) or correct (1 point) and the individual total scores as well as the mean scores were calculated for each of the 36 participants who completed the test³. On average, the participants scored 16.7 out of 25 possible points ($SD = 4.17$), with a minimum of 7 points and a maximum of 23 points. The distributions have a skewness of $-.65$ and kurtosis of 2.69 which implies a skewed to the left distribution of test scores. Although none

3 Not all students completed all three tests: The results of 8 students are missing for the intelligence test and of 19 students for the WiWiKom test.

of the participants achieved a maximum score, an examination of the answers at item level indicates that the participants were able to complete the test in the given time. In comparison, a nationwide German sample of 7,571 bachelor's students of economics in the WiWiKom study achieved an average score of 13.3 points ($SD = 4.39$; Zlatkin-Troitschanskaia et al. 2019a). The students in our test sample distinguish themselves from this latter sample by a comparatively high level of economics knowledge.

Using a Spearman test, we found no significant correlation between the PAL scores and the WiWiKom test scores ($r_s = .11$, $p = 51.2$, $n = 36$). The correlation also remains insignificant if calculated only for economics students ($r_s = -.04$, $p = 0.85$, $n = 28$).

For further analyses, the relationship between the PAL task and economic knowledge was examined by means of cross-classified tables which investigated the interrelations of different performance levels in both tests. For this purpose, using the median of the PAL scores and of the WiWiKom test, the sample was divided up into equally large groups of high and low performers. Table 4 shows the cross-classification of overall performance in PAL and in the WiWiKom test. For the cross-classified tables of the subscales, the groups were clustered according to the score of the respective subscale in the PAL task.

Table 4 Cross-classification of overall performance in PAL and the WiWiKom test

Performance in PAL	Performance in WiWiKom test		
	Low (=0)	High (=1)	Total
Low	7 46.67 %	8 53.33 %	15 100.00 %
High	8 38.10 %	13 61.90 %	21 100.00 %
Total	15 41.67 %	21 58.33 %	36 100.00 %

The cross-classification shows that 47 % of students who performed low on the PAL task are also low-performers on the WiWiKom test. However, the majority (53 %) of low performers comprises students who performed well in the WiWiKom test. Yet again, it becomes apparent that despite high levels of domain-specific knowledge, overall, economics students did not perform particularly well in the PAL task. This further supports the hypothesis (H2) that many of these students were not able to apply their domain-specific knowledge to solve the economics problem.

Conversely, 38 % of the students with high performance in the PAL task show low performance in the WiWiKom test. Over one third of students achieved good results in the PAL task despite low levels of domain-specific knowledge, which they evidently were able to compensate with higher general cognitive abilities (probably by reasoning and adequate use of the given information). A majority (62 %) performed well in both tests. The chi-squared test is not significant for the cross-classified table ($\chi^2(36) = 0.26, p = .607$), which leads to the overall conclusion that high economics knowledge does not necessarily indicate a high performance in the PAL task.

Similar findings were derived from the cross-classified table for the subdimension "Recognizing and evaluating the relevance, reliability, and validity of given information" ($\chi^2(36) = 0.39, p = .53$), in which, on average, participants achieved the highest scores. The findings for the subdimension "Evaluating and making a decision" ($\chi^2(36) = 1.03, p = .31$) showed that two thirds of high performers in the WiWiKom test performed poorly in the PAL task. Based on these results, *H2*, assuming a positive but relatively weak correlation between the critical thinking performance of students and their level of economic knowledge, must be rejected.

As the PAL task is constructed to assess higher cognitive skills, a slight correlation with the general cognitive abilities measured in the intelligence test was expected (*H3*).

The results of the 20 intelligence test items in each of the two employed scales were calculated (0 = wrong answer, 1 = right answer) and participants could achieve a maximum of 40 points in total on the test. On average, the students achieved a score of 16.9 ($SD = 5.33$) with a minimum of 7 and a maximum of 29 points. Because of skewness to the left of the PAL results, a Spearman correlation was conducted. As expected, the results show a significant weak correlation between the PAL scores and the intelligence test scores ($r_s = .35, p = .02, n = 46$).

With regard to the students' school-leaving grades, there were no related differences in performance in the PAL task; students with a final grade of 2.0–2.9 performed better than students with either a final grade of 1.0 – 1.9 or 3.0 and higher. Table 5 shows the means of PAL scores subdivided by different groups. The Spearman correlation of these two performance measures indicates no significant correlation ($r_s = -.03, p = .82, n = 47$). Based on these findings, *H3* cannot be rejected.

To test for the group differences assumed in *Hypothesis 4* (*H4*) (Table 5), we conducted a t-test which shows no significant differences between students of economics and students with other majors with regard to average test scores ($p = 0.69$). Thus, contrary to our expectations (*H4*), studying economics does not appear to

provide domain-specific knowledge that enhances the ability to solve these tasks.⁴ Based on these results, *H4* must be rejected.

Table 5 Means of PAL-scores of different groups

	Group Average Performance in PAL
Variables	<i>N</i> = 55
Degree	
Bachelor	3.53 (SD = .73)
Master	3.68 (SD = .67)
Subject	
Economics studies	3.62 (SD = .68)
Other	3.52 (SD = .82)
University entry qualification grade	
1.0 – 1.9	3.62 (SD = .85)
2.0 – 2.9	3.66 (SD = .60)
3.0 – 3.9	3.55 (SD = .69)
Completed Vocational Training	
No	3.58 (SD = .79)
Yes (commercial)	3.69 (SD = .59)

5 Discussion and Conclusion

This study primarily presented findings on performance-oriented assessment of key facets of domain-specific and generic skills such as critical thinking and content knowledge among economics students at a German university. In this validation study, the approach of comparing known groups by assessing undergraduate economics students as well as a control group of master's students in economics degree programs has been applied. Additionally, for a domain-specific comparison, a control group of students with different major subjects was assessed in comparison to students having economics as their main study subject.⁵ Beside the main finding that the construct of 'critical thinking' measured by the PAL task has turned out to be of discriminant validity, the results provide two important

- 4 The comparison between economics students ($n=45$, 12 bachelor's and 33 master's students) and students with other main subjects ($n=8$) shows that, on average, economics students achieve slightly better test results (economics: 3.6 points, others: 3.5 points).
- 5 The descriptive statistics of the sample are largely in line with the results of a Germany-wide survey of graduates of economics education at 20 universities, which further increases the external validity of the results of this study (see Kuhn et al. in this volume).

findings: (1) bachelor's students performed rather poorly in the PAL task on average and the level of critical thinking in master's students is much lower than one would expect of graduate students at the end of their 5-year academic education. (2) There were no significant differences between economics students and students with other main study subjects.

This study provides several pieces of evidence that many students in economics are not able to apply their domain-specific knowledge to solve realistic economics-related problems. This finding appears even more dramatic in light of performance in each of the four subdimensions. Overall test performance, which on average is mediocre, is primarily based on well-developed abilities in rather basic (1) information processing and (4) writing (Table 3). With respect to the two other subdimensions of critical thinking that were considered vital in the construct definition (i.e. (2) decision making and (3) dealing intellectually with consequences), most students displayed substantial deficits – even towards the end of their academic education, and also at the graduate level.

Overall, the findings indicate that although critical thinking skills are required both in curricula and as learning outcomes in economic higher education, these key aspects have been insufficiently or ineffectively nurtured so far, both in undergraduate and in graduate studies. Although students have a relatively solid level of economic knowledge and understanding, they clearly lack the ability to transfer this knowledge and to solve a concrete economics-related problem. Based on the results from the innovative performance assessment, it is urgently necessary to give careful consideration to why the highly ambitious teaching-and-learning objectives and the expected outcomes of higher education are possibly not reached at all.

Our results raise a number of questions, which require more in-depth analyses to gain differentiated insights as to how such higher cognitive skills can be effectively taught and enhanced during higher education in economics. Multiple-choice tests are, at least in Germany, usually used for examinations, while tests based on a case study are rare (e.g., Walstad 2001). Therefore, a testing effect cannot be ruled out; it is possible that, although students do gain domain-specific knowledge, they cannot demonstrate it in a PAL test due to the unfamiliarity with this type of test instrument. A controlled, experimental intervention study should be conducted with a posttest measurement design to measure such effects.

Additionally, this validation study was conducted using a small sample from a single university. Thus, the results presented here should be considered only as preliminary evidence for the level of critical thinking measured by this particular PAL task and they still leave room to expect better results based on an ample sample. Furthermore, in future studies, not only performance data, but also behavior

data and students' response processes while working on PAL tasks, such as log and gaze data, should be collected and integrated in analyses to assess which concrete cognitive and non-cognitive subskills are used to solve performance tasks.

References

- Alexander, P. A., Singer, L. M., Jablansky, S., & Hattan, C. (2016). Relational reasoning in word and in figure. *Journal of Educational Psychology*, 108 (8), 1140–1152.
- American Educational Research Association, American Psychological Association and National Council on Measurement in Education (AERA, APA and NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Allgood, S., & Bayer, A. (2016). Measuring College Learning in Economics. MPR Paper No. 85104. Online: <https://mpr.ub.uni-muenchen.de/85104/>.
- Allgood, S., Walstad, W. B., & Siegfried, J. J. (2015). Research on Teaching Economics to Undergraduates. *Journal of Economic Literature*, 53 (2), 285–325.
- Arum, R., & Roksa, J. (2010). Academically adrift: Limited learning on college campuses. Chicago, IL: University of Chicago Press.
- Browne, M. N., Hoag, J. H., & Boudreau, N. (1995). Critical Thinking in Graduate Economic Programs: A Study of Faculty Perceptions. *The Journal of Economic Education*, 26 (2), 177–181.
- Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relation to critical thinking skill. *Informal Logic*, 20 (1), 61–84.
- Fives, H., & Dinsmore, D. L. (Eds.). (2017). *The Model of Domain Learning: Understanding the Development of Expertise*. New York, Abingdon: Routledge.
- Halpern, D. F. (2014). *Thought and knowledge: An introduction to Critical Thinking*. New York: Psychology Press.
- Happ, R., Nagel, M.-T., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2019). How migration background affects master degree students' knowledge of business and economics. *Studies in Higher Education*. doi: 10.1080/03075079.2019.1640670.
- Hattie, J., & Cooksey, R. W. (1984). Procedures for assessing the validities of tests using the "known-groups" method. *Applied Psychological Measurement*, 8 (3), 295–305.
- Hoyt G. M., & McGoldrick, K. (Eds.). (2012). *International Handbook on Teaching and Learning Economics*. Cheltenham, Northampton: Edward Elgar.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
- Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). Validity of the SAT® for Predicting First-Year College Grade Point Average. *College Board Research Report* (5). New York: The College Board.
- Lai, E. R., & Viering, M. (2012). Assessing 21st Century Skills: Integrating research findings. Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, B.C., Canada.
- Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Available on <https://psychowissen.jimdo.com/psychologisches-tests/intelligenztests/i-s-t-2000-r/>

- Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessments. *ETS Research Report*, 14 (10).
- McGoldrick, K., & Garnett, R. (2013). Big Think: A Model for Critical Inquiry in Economics Courses. *The Journal of Economic Education*, 44 (4), 389–398.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement (3rd ed., pp. 13–103)*. New York: Macmillan Publishing.
- Paul, R. W., & Elder, L. (2006). Critical thinking: The nature of critical and creative thought. *Journal of Developmental Education*, 30 (2), 34–35.
- Pellegrino, J. W., & Hilton, M. L. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National Academies Press.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. (2019). Assessment of University Students' Critical Thinking: Next Generation Performance Assessment. *International Journal of Testing*. doi: doi.org/10.1080/15305058.2018.1543309
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Mariño, J. P. (2018). *International Performance Assessment of Learning in higher education (iPAL) – Research and Development*. Wiesbaden: Springer.
- Walstad, W. B. (2001). Improving Assessment in University Economics. *The Journal of Economic Education*, 32 (3), 281–294.
- Walstad, W. B., Rebeck, K., & Butters, R. B. (2013). *Test of economic literacy: Examiner's manual*, 4th ed. New York: Council for Economic Education.
- Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of Understanding in College Economics: Examiner's manual*, 4th ed. New York: National Council on Economic Education.
- West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, 100 (4), 930–941.
- Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? *American Educator*, 31 (2), 8–19.
- Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., Förster, M. & Brückner, S. (2019). Validating a Test for Measuring Knowledge and Understanding of Economics Among University Students. *Zeitschrift für Pädagogische Psychologie*, 33(2), 119–133.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019b). On the complementarity of holistic and analytic approaches to performance assessment scoring. *The British Journal of Educational Psychology*. Advance online publication. <https://doi.org/10.1111/bjep.12286>
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Pant, H. A. (2018a). Assessment of Learning Outcomes in Higher Education – International Comparisons and Perspectives. In C. Secolsky and B. Denison (Eds.). *Handbook on Measurement, Assessment and Evaluation in Higher Education (2nd ed.)*. New York: Routledge.
- Zlatkin-Troitschanskaia, O., Toepfer, M., Molerov, D., Buske, R., Brückner, S., Pant, H. A., Hofmann, S., & Hansen-Schirra, S. (2018b). Adapting and Validating the Collegiate

Learning Assessment to Measure Generic Academic Skills of Students in Germany: Implications for International Assessment Studies in Higher Education. In O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, C. Kuhn (Eds.) *Assessment of Learning Outcomes in higher education – Cross-National Comparisons and Perspectives* (pp. 245–266). Cham: Springer.



3.7

The Research Group Performance-Based Assessment of Communication in KoKoHs

A Bridge between Educational Theory and Empirical Educational Research

Falkenstern, A., Schwabe, U., Walz, K., and Braun, E.

Abstract

In recent years, educational theory and empirically oriented educational research have been viewed in contrary light. Growing interest in evidence-based expertise in education and the increase in assessment of competences reinforced this debate. The aim of this contribution is to question this contradiction, and to consider possibilities of constructive collaboration between educational theory and empirical research in assessment of competences. The aim is to stimulate new formats of assessment. We discuss education as a holistic transformation and a social process. According to this understanding, we identify communication competences as an educational aim and discuss opportunities of its training and assessment. We suggest role-plays as an adequate method to develop a competence-based instrument. Ten specific role-plays were developed, which were applied by $N = 515$ students. The results support role-plays as an effective and reliable assessment of communication competences. Nevertheless, further analyses are necessary to detect and explain various effects and gains in competences. The combination of empirical educational research and educational theory seems to be promising, and we hope to discover further conceptual impacts by bringing both sides together.

Keywords

Performance based test, communication competences, educational theory, social interaction, higher education, intersubjectivity, KoKoHs

1 Theoretical Considerations

1.1 Empirical Turn in Educational Research and Critical Discourse

In recent years, there has been a growing interest in scientific and evidence-based expertise for education policy. Empirical educational research provides data and information in purpose of educational monitoring and comparative studies, which are used for practical and educational policy decisions (Bromme et al. 2016). This trend also affects higher education, as research in the measurement of competences in higher education has increased substantially (Coates and Zlatkin-Troitschanskaia 2019; OECD 2012; Zlatkin-Troitschanskaia et al. 2017).

At the same time, critical concerns¹ from theoretically oriented educational research arise. There is critique about making “everything”, including education, measurable and standardizable, in doubt of comprising education to any statistical values. Theoretical education and empirical research on competences seem to be on two sides of research. One frequently expressed critique concerns measuring practices, characterized by analytical rationality and evidence orientation, which are incompatible with the classical identity of education (“Bildung”). From the point of view of educational theory (“Bildungstheorie”), current empirical education and competence research rarely deals with classical understanding of education and fails to address their subject. Education is reduced to the acquisition of certain formal and domain-specific skills, therefore the current knowledge on education is superficial (e.g. Schäfer 2006; Lederer 2014; Hastedt 2012; Pongratz et al. 2007; Klein and Dungs 2010).

Since the expansion of empirical educational research (Gräsel 2011), philosophical concepts of education seem to become obsolete and out-of-date categories (Koller 2010, p. 93f.; Koller 2012, p. 10; Ehrenspeck 2010; Lederer 2014, p. 69ff.). Generally, modern empirically educational research does not focus on philosophically defined educational ideals and concepts (Gräsel 2011), but on the societal

¹ Baumert and Tillmann (2016) provide a broad overview of the current critical discussions about empirical educational research.

function of education and on empirically measurable learning outcomes. Modern empirical educational research claims to follow the positivist scientific ideal and to present itself as a modern and evidence-based discipline along quality criteria such as standardization, operationalization and comparability (e.g. Reinders and Ditton 2011).

In recent years, both sides have often been viewed in a contrary light dealing with tensions. This relationship is usually regarded as incompatible (Fuchs 2011, p. 31ff.). Classical educational theories based on philosophy of education are looking at holistic educational processes and therefore not standardizable situations (Koller 2010). At the same time, educational theory blames empirical educational research and competence research of not measuring what they actually supposed to measure and of not being able to capture meaningful educational processes.

The aim of this contribution is to question this opposition, to consider productive interactions of educational theory and empirical competence research. The consideration of the relation between educational theory and competence research promises fruitful results and the possibility to develop suitable instruments for assessing gains of competencies in higher education by investigating effective learning opportunities. In addition, it will be shown that educational processes can be captured with certain methods and that individual aims educational processes can be made empirically visible. This contribution aim to demonstrate that a performance-based instrument of competence assessment presented here is compatible with educational theory, since this method is based on a holistic understanding of education.

1.2 Education as a Holistic Transformation Process

Following the theory of education, education is more than accumulation of factual knowledge (e.g. Marotzki 1997). Education represents a holistic transformation of a person (e.g. Koller 2012; Nohl 2006; Fuchs 2011; Marotzki 1990; von Rosenberg 2011), who develops in the interaction with the environment and in self-interaction. Education can be considered as a dynamic interaction between subject and environment (Koller 2012). This understanding of education is based on a subject-oriented approach and focuses on the dynamics of individual educational processes. However, empirical educational research often reduces educational outcomes, the empirical connectivity of holistic transformation seems hardly achievable (Tillmann 2017; Pant 2013).

There is an ongoing discourse on generic competences in higher education (Braun et al. 2018a; Braun et al. 2018b). The discussion implies that education is

not limited to the acquisition of factual knowledge and information reception, but also includes support of generic skills and a holistic understanding of education. In higher education, students are not supposed to be filled up with domain-specific knowledge and information, analogous to a ‘*tabula rasa*’. In addition to impart knowledge, higher education institutions are responsible for promoting self-development and taking responsibility. Academic education also aims to prepare students for the ever-changing and unpredictable demands of the world and the labour market. The German Science Council (Wissenschaftsrat 2015) emphasises that the academic educational goals should focus on three central dimensions: specialised knowledge, self-development and preparation for the labour market.

In recourse to the classical educational ideals², independent thinking and autonomous action are important outcomes of educational processes. The ability to use one’s own intellect (without the guidance of others) plays an increasingly important role in times of information society. In view of the technical edge of information and communication technologies, information is continuously increasing and provides unlimited access to many resources of knowledge. The knowledge of the total world and the existing information increases with rapid speed (e.g., Lederer 2014, p. 230). These developments affect higher education as well. The quantitative increase in knowledge can create uncertainties with regard to options for knowledge and action. Skills are required to be able to orientate oneself and to be able to act competent in unknown situations. The knowledge application and the ability to use the accumulated knowledge play a central role in today’s information-based society.

1.3 Education as a Social Process

By referring to the theoretical reflections on education, educational processes are reliant on social and societal context, since the subject is embedded inevitably in social contexts. Learning processes cannot occur isolated from the social counterpart and do not emerge automatically from oneself, but become possible in the active confrontation and interaction between subject and its environment (Grunert 2012, p. 35ff.; Vygotsky 1978). Individual learning processes and the formation of self are caused by social and communication processes; individual ability to act is not a one-sided result of individual learning effort and self-education, but can be acquired in interplay between environment and subject (Grunert 2012, p. 35ff.). In

2 Dörpinghaus et al. (2012), Ladenthin (2012) provide, for example, an overview of classical educational theories based on educational philosophy.

this respect, educational processes can be considered as a dialectical process: individuals' learning processes are embedded in intersubjective and social experience contexts (Mead 1973; Vygotsky 1978). Therefore, also gain in competences can be seen as a product of social and interactive processes. Due to the social intermediation of educational processes, the ability to act appropriately at given social contexts is significant to master a situation. Educational processes are not the result of isolated mental processes, but are embedded in communicative actions and experiences (Lederer 2014, p. 154f.). The ability to communicate effectively and interact with the environment is essential to master nowadays requirements. The ability to act cannot be acquired as factual knowledge and is supported through practical action and experiences. Learning processes emerge in the direct confrontation with the environment and in the self-encounter. Educational processes presuppose intersubjectivity as individuals depend on their social counterparts (Grunert 2012, p. 27). Self-formation can hardly develop without the sociality. Subject formation occurs during social and communicative interactions (Mead 1973; Vygotsky 1978). Individual holistic educational processes are therefore only conceivable in social terms, and the ability to interact and communicate with the environment is important.

1.4 Communication Competence as an Educational Aim

According to Habermas, communication is essential for coordinating human action and human communities (Habermas 1999a; 1999b; Schneider 2009; Wesseling 2012). Communicative action requires specific abilities (Habermas 1984; 1971). Communication is based on patterns of orientation and legitimation, which are socially conveyed and internalised by the acting person. Successful communication is related to the interpretative abilities and to the norm-based understanding. This includes the knowledge of socially accepted standards and communication strategies. Communicative competences include knowledge of the communication rules and of the socially shared patterns of interpretation and can be defined as the ability to adequately apply social rules in communicative actions (Habermas 1999a; Schneider 2009, p. 212ff.).

For example, knowledge refers to how to communicate appropriately in various situations with individuals in different social roles, and to the ability to use language and elements of nonverbal communication appropriately. The ability to adequately interpret the communication situation and to know the rules of interaction and strategies for efficient achievement of goals is significant for successful interaction (Braun et al. 2018a; Knoblauch 2010, p. 245f.). The ability to differen-

tiate, how to communicate with interaction partners of diverse social roles, how to effectively achieve goals in communication and under what conditions communication can fail, can be described as communicative capability (Braun et al. 2018a; Frindte 2001, p. 35ff.). Habermas (1999a) differentiates between *strategic action* as a form of success-oriented action and communicative action as *consensus-oriented action*.³

Successful communication requires knowledge of situational and social factors. Depending on communication goals and the context-specific, given conditions, different communication strategies can be used, assuming that the interaction partners know their social role, limitations and legitimacy. The possibilities for action can be based on this knowledge. Communicative competence also includes the ability to achieve goals that are linked to communication without massively damaging the interests of another person (Röhner and Schütz 2016).

Habermas' theory of communicative action also considers individual's ability to learn. Social and living environment become accessible only when individual enter into interpersonal interactions and use their communicative abilities (Weseling 2012). Habermas focuses not only on the cognitive aspects, but also on the individual's capability to act and communicate, and also uses the term "interaction competence" (Habermas 1984; 1974).

1.5 Training and Assessing Communication Competences

Besides providing formal and domain specific knowledge, one central task of higher education can be considered in imparting generic competences. Especially, the so-called Bologna Process emphasised the importance of generic competences and of clarifying learning outcomes. Communication is considered a central learning outcome in higher education, and therefore explicitly defined as an educational objective (Kultusministerkonferenz 2017; European Commission 2008). Competences are defined as the ability to solve problems in different situations, which can be described as the ability to act sensitive of the given situation (Grunert 2012, p. 60f.). In addition, the application of knowledge is emphasised. If knowledge is acquired through complex exercises, the transfer of knowledge to new and authentic situations will be easier. Furthermore, professional action plays an important role in higher education. The professional fields, graduates will have to master, cannot be standardized. Therefore, specialist knowledge is not sufficient. Graduates should be able to apply the acquired knowledge in a goal-oriented and

3 For the implementation of communication types in role-plays, see Braun et al. (2018b).

context-appropriate way. It is hardly possible to draw up a homogeneous view of practice and to provide students with clearly defined instructions for action in the sense of “know-how”.

One objective of higher education is to enable students to solve problems and conflicts. Professional action means being able to decide which form of action is appropriate in a specific situation. Professional practice is characterised by the aspect of uncertainty and ambivalence (Cramer and Drahm 2019, p. 21; Cramer et al. 2019, p. 24). Due to the complexity of practice, it is hardly possible to provide universally valid, standardised strategies for action. Professional action refers to both the acquired professional knowledge, which can be defined as formal knowledge, and practical knowledge and knowledge in action, which are also essential for professional action (Cramer and Drahm 2019, p. 24). Professionals are characterized by the ability to successfully cope with unknown and complex situations and to interact with the environment without major conflicts. Practical knowledge and competences such as communication skills can be acquired best by experience in authentic situations and practical demands. Higher education can provide learning opportunities in which students can acquire their skills in an application-oriented way. Furthermore, students could be stimulated to master complex situations and situation-specific demands, if they are in challenging learning environments.

1.6 Performance-Based Instrument

Summing up, the question arises how learning opportunities for professional action can be designed in higher education. Role-plays seem to be an adequate performance-based instrument to initiate communication and to provoke acting in complex and authentic situation (Braun and Mishra 2016; Gulikers et al. 2008; Stevens 2015). Role-plays can be used as a method of professional training in which specific, significant, and professional situations are simulated. In this way, students can train and show the ability to act in authentic situations. Role-plays can be adapted to several specific professional contexts and still are conferrable. The situational link to action and the moments of uncertainty are manifest in role-plays. Role-plays as a method create conditions for coping with complex situations and challenges (Braun et al. 2018a, 2018b).

Habermas referred to the concept of performance in his theory of communicative action. The concept of performance plays a central role in theories of speech act (Habermas 1988), since social action and communication are language-mediated processes. Performance is activated through communicative situations, in which social aspects play a fundamental role and action is needed (Knoblauch

2010, p. 242f.). As mentioned, educational processes are embedded in interactive processes. Role-plays demand an active interaction with social counterpart. Since competences mean also an efficient use of knowledge, competences are interconnected with performance. By playing different roles and scenarios, the ability of changing perspectives can be developed. Role-plays might also initiate processes of reflection: the content of role-plays is designed to provoke autonomous thinking and action. Students cannot apply predefined solutions – because of the complex and unknown situation – but are encouraged to think autonomously and make situational decisions (Braun and Mishra 2016; Gulikers et al. 2008; Stevens 2015).

Such an instrument is compatible with the holistic understanding of education, since education is not reduced to the accumulation of formal knowledge. Educational processes are viewed in an expanded, holistic understanding grounded in social communication and interactional contexts. Generic competences such as communicative competences can also be described as person-specific skills, since these skills are performed and developed individually. Interaction with the environment is constitutive for educational processes. Following this assumption, role-plays can create a link to social aspects of action. In addition, authentic and complex situations allow observation and therefore empirical assessment. Following these theoretical considerations, we have developed a performance-based instrument that provides learning environments for communicative skills and makes transformations in learning processes possible and visible.

To highlight the connectivity of empirical educational research and educational theory (“Bildungstheorie”), conceptually and empirically, we discuss some empirical findings of the developed performance-based role-plays in the light of educational theory. Our assumption is that the developed performance-based instrument is reliable in the sense of test-theory. In this chapter, we focus on two questions for empirical examination:

- a) How internal consistent are the role-plays of the performance-based instrument?
- b) How stable is the behaviour estimated in the performance-based test over time?

2 Method

2.1 Test Setting

A research group, based on the KoKoHs project, developed ten role-plays to simulate authentic, complex situations, which require social action. Each role-play con-

tains a short instruction, describing an uncertain and ambivalent situation. There are five role-plays for each orientation of action; either *success-oriented action* or *consensus-oriented action* (Habermas 1999a). It is up to the student to decide how to perform and how to communicate in the situation. Observers assess the performance using a standardized observation form based on theoretical considerations. All role-plays contain 9 to 16 items (for details, see Braun et al. in revision).

There are three people needed to fulfill the assessment situation: First, the student whose communication competences is assessed. Second, the trained conversation partner. Third, the trained observer/rater. Training and observation form ensure a certain level of standardization, while authenticity in conversation style induced a design-restricted openness.

2.2 Sample

A first empirical implementation of the role-play assessment was taking place in 11 different German higher education institutions (six universities, four universities of applied sciences, one private college), where 515 students participated in the role-plays. The higher education institutions have been a random sample of all German colleges, and students participated voluntarily. Each participant performed four out of ten role-plays. Due to this design, there are 40 % observed values and 60 % values missing at random. The sequence of the role-plays was randomized.

The participants from two higher education institutions, $N = 52$ students, performed the role-plays twice, with two weeks in between. This second measurement serves for the check of retest-reliability. Due to resource constraints and different places of the institutions all over Germany, it was not possible to assess all participants two times.

2.3 Analysis of Reliability

We analyzed two different forms of reliability. In a first step, we estimated *Cronbach's alpha* to test the consistence of the measured construct. The higher the value of *Cronbach's alpha*, the better observed items measure one latent construct (internal consistency). For a further reliability check, we used the subsample with two measurements to test the retest reliability and to analyze the stability of the observed construct, the communication competences. Pearson correlations for each communication action are calculated to estimate the stability of communication competences.

3 Results

In a first step, we calculated the *Cronbach's alpha* for both communication styles. The results show high alpha-values, for *strategic communication* alpha is .96, and for *consensus-oriented communication* alpha is also .96. The observers were able to use the standardized observation forms to assess students' ability in a consistent manner.

In a second step, we are interested in retest-reliability. If communication competences should be measurable, the stability of the competence is one precondition – in the sense that someone is competent or less competent over the time. Communication competence is considered learnable and therefore modifiable.

The analyses conducted medium and significant coefficients for the retest reliability of the dimension strategic communication ($r = 0.46, N = 52, p < .01$) and for the consensus-oriented ($r = 0.27, N = 52; p < .05$) communication.

4 Discussion

Summarizing the results of the reliability analyses, observers were able to rate *strategic communication* and *consensus-oriented communication* in a consistent manner, and therefore, the test is reliable in the sense of internal consistence. The results of the retest reliability are more heterogeneous. We observed an interesting reaction of the students. After the second time of the assessment, some of the students' report, that they have explicitly used another communication-orientation after the experience of the first time. They report the effect of the first time has been impressive, and they thought a lot about the role-plays afterwards. They experience the role-plays as meaningful learning opportunity, and some of the participants reported to felt irritated after the session. Therefore, we assume the performance-based instrument follows the considerations of creating an environment, which allows complex social and in self-interactions. Role-play has the opportunity to provoke autonomous thinking and action, and reflection. Therefore, the retest reliability may have some restriction as an evidence-based criterion for performance-based tests.

It has to be mentioned the limitations of our contribution. So far, we have used some easy-accessible criteria, such as Cronbach and Pearson correlations. However, deeper analysis and a larger sample are necessary to explain some 'non-stability' of our re-test. How and why has the behaviour changed, when is it stable? Are the students more irritated or have they gained competences between both times?

The connection of empirical educational research and educational theory seems to be promising, and we expect to discover further conceptual impacts by bringing both sides together. We hope to contribute to the question of Shavelson et al. (2018), how to measure complex and holistic educational outcomes.

References

- Baumert, J., & Tillmann, K.-J. (Eds.) (2016). Empirische Bildungsforschung. Der kritische Blick und die Antwort auf die Kritiker. *Zeitschrift für Erziehungswissenschaft*, 31. Wiesbaden: Springer Fachmedien.
- Braun, E., & Mishra, S. (2016). Methods for assessing competences in higher education: A comparative review. In *Theory and method in higher education research* (pp. 47–68). Emerald Group Publishing Limited.
- Braun, E., U. Schwabe, & Klein, D. (in revision). Kompetenzorientierte Prüfung kommunikativer Fähigkeiten von Studierenden. *Diagnostica*.
- Braun, E., Athanassiou G., Pollerhof, K., & Schwabe, U. (2018a). Wie lassen sich kommunikative Kompetenzen messen? – Konzeption einer kompetenzorientierten Prüfung kommunikativer Fähigkeiten. *Beiträge zur Hochschulforschung*, 40 (3), (pp. 34–55).
- Braun, E., Schwabe, U., & Klein, D. (2018b). Performance-based tests: using role plays to assess communication skills. In S. McGrath, M. Mulder, J. Papier, & R. Stuart (Eds.), *Handbook of vocational education and training: developments in the changing world of work* (pp. 1–11). Cham: Springer International Publishing.
- Bromme, R., Prenzel, M., & Jäger, M. (2016). Empirische Bildungsforschung und evidenzbasierte Bildungspolitik. Zum Zusammenhang von Wissenschaftskommunikation und Evidenzbasierung in der Bildungsforschung. *Zeitschrift für Erziehungswissenschaft*, 19 (1), (pp. 129–146).
- Coates, H., & Zlatkin-Troitschanskaia, O. (2019). The Governance, Policy and Strategy of Learning Outcomes Assessment. *Higher Education Policy*, Special Issue, (pp. 1–6).
- Cramer, C., & Drahmman, M. (2019). Professionalität als Meta-Reflexivität. In M. Syring, & S. Weiß (Eds.), *Lehrer(in) sein – Lehrer(in) werden – die Profession professionalisieren* (pp. 17–33). Bad Heilbrunn: Klinkhardt.
- Cramer, C., Harant, M., Merk, S., Drahmman, M., & Emmerich, M. (2019). Meta-Reflexivität und Professionalität im Lehrerinnen- und Lehrerberuf. *Zeitschrift für Pädagogik*, 65(3), (pp. 401–423).
- Dörpinghaus, A., Poentisch, A., & Wigger, L. (2012). *Einführung in die Theorie der Bildung*. Darmstadt: WBG.
- Ehrenspeck, Y. (2010). Philosophische Bildungsforschung: Bildungstheorie. In R. Tippelt & B. Schmidt (Eds.), *Handbuch Bildungsforschung* (pp. 155–170). Wiesbaden: VS Verlag.
- European Commission. (2008). *The European Qualifications Framework for Life Long Learning* (EQF). European qualifications framework. http://www.ecompetences.eu/site/objects/download/4550_EQFbroch2008en.pdf. Accessed: 01 July 2019.
- Frindte, W. (2001). *Einführung in die Kommunikationspsychologie*. Weinheim/Basel: Beltz Verlag.

- Fuchs, T. (2011). *Bildung und Biographie: Eine Reformulierung der bildungstheoretisch orientierten Biographieforschung*. Bielefeld: Transcript.
- Gräsel, C. (2011). Was ist Empirische Bildungsforschung? In H. Reinders, H. Ditton, C. Gräsel & B. Gnieuwosz (Eds.), *Empirische Bildungsforschung. Strukturen und Methoden* (pp. 13–28). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Grunert, C. (2012). *Bildung und Kompetenz. Theoretische und empirische Perspektiven auf außerschulische Handlungsfelder*. Wiesbaden: Springer VS.
- Gulikers, J. T., Kester, L., Kirschner, P. A., & Bastiaens, T. J. (2008). The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes. *Learning and Instruction, 18*(2), (pp. 172–186).
- Habermas, J. (1999a). *Theorie des kommunikativen Handelns*. Bd. 1. Handlungsrationalität und gesellschaftliche Rationalisierung. Frankfurt am Main: Suhrkamp.
- Habermas, J. (1999b). *Theorie des kommunikativen Handelns*. Bd. 2. Zur Kritik der funktionalistischen Vernunft. Frankfurt am Main: Suhrkamp.
- Habermas, J. (1988). Handlungen, Sprechakte, sprachlich vermittelte Interaktionen und Lebenswelt. In J. Habermas (Ed.), *Nachmetaphysisches Denken. Philosophische Aufsätze* (pp. 63–104). Frankfurt am Main: Suhrkamp.
- Habermas, J. (1984). Notizen zur Entwicklung der Interaktionskompetenz. In J. Habermas (Ed.), *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns* (pp. 187–225). Frankfurt am Main: Suhrkamp.
- Habermas, J. (1974). Moralentwicklung und Ich-Identität. In J. Habermas (Ed.), *Zur Rekonstruktion des Historischen Materialismus* (pp. 63–91). Frankfurt am Main: Suhrkamp.
- Habermas, J. (1971). Vorbereitende Bemerkungen zu einer Theorie der kommunikativen Kompetenz. In J. Habermas & N. Luhmann (Eds.), *Theorie der Gesellschaft oder Sozialtechnologie – Was leistet die Systemforschung?* (pp. 101–141). Frankfurt am Main: Suhrkamp.
- Hastedt, H. (2012). *Was ist Bildung? Eine Textanthologie*. Stuttgart: Reclam.
- Klein, R., & Dungs, S., (Eds.). (2010). *Standardisierung der Bildung. Zwischen Subjekt und Kultur*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Knoblauch, H. (2010). Von der Kompetenz zur Performanz. Wissenssoziologische Aspekte der Kompetenz. In Th. Kurtz & M. Pfadenhauer (Eds.), *Soziologie der Kompetenz* (pp. 237–256). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Koller, H.-C. (2017). *Grundbegriffe, Theorien und Methoden der Erziehungswissenschaft. Eine Einführung*. 8., aktualisierte Auflage. Stuttgart: Kohlhammer.
- Koller, H.-C. (2012). *Bildung anders denken. Einführung in die Theorie transformatorischer Bildungsprozesse*. Stuttgart: Kohlhammer.
- Kultusministerkonferenz. (2017). *Qualifikationsrahmen für Deutsche Hochschulabschlüsse*. KMK. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2017/2017_02_16-Qualifikationsrahmen.pdf. Accessed: 01 July 2019.
- Ladenthin, V. (Ed.). (2012). *Philosophie der Bildung. Eine Zeitreise von den Vorsokratikern bis zur Postmoderne*. Bonn: DenkMal Verlag.
- Lederer, B. (2014). *Kompetenz: oder Bildung. Eine Analyse jüngerer Konnotationsverschiebungen des Bildungsbegriffs und Plädoyer für eine Rück- und Neubesinnung auf ein transinstrumentelles Bildungsverständnis*. (Habilitationsschrift, Universität Innsbruck.). Innsbruck: Innsbruck university press.

- Marotzki, W. (1997). Morphologie eines Bildungsprozesses. Eine mikrologische Studie. In D. Nittel, & W. Marotzki (Eds.), *Berufslaufbahn und biographische Lernstrategien. Grundlagen der Berufs- und Erwachsenenbildung*, Bd. 6 (pp. 83–117). Baltmannsweiler: Schneider Verlag Hohengehren.
- Marotzki, W. (1990). *Entwurf einer strukturalen Bildungstheorie. Biographietheoretische Auslegung von Bildungsprozessen in hochkomplexen Gesellschaften*. Weinheim: Deutscher Studien Verlag.
- Mead, G. H. (1973). *Geist, Identität und Gesellschaft aus der Sicht des Sozialbehaviorismus*. Frankfurt am Main: Suhrkamp.
- Nohl, A.-M. (2006). *Bildung und Spontaneität. Phasen biographischer Wandlungsprozesse in drei Lebensaltern. Empirische Rekonstruktionen und pragmatische Reflexionen*. Oppladen: Barbara Budrich.
- OECD. (2012). *Bessere Kompetenzen, bessere Arbeitsplätze, ein besseres Leben: Ein strategisches Konzept für die Kompetenzpolitik*. OECD Publishing. https://www.oecd-ilibrary.org/education/bessere-kompetenzen-bessere-arbeitsplatze-ein-besseres-leben_9789264179479-de. Accessed: 10 July 2019.
- Pant, H. A. (2013). Wer hat einen Nutzen von Kompetenzmodellen?. *Zeitschrift für Erziehungswissenschaft*, *16*(1), 71–79.
- Pongratz, L. A., Reichenbach, R., & Wimmer, M. (Eds.). (2007). *Bildung – Wissen – Kompetenz*. Bielefeld: Janus Presse.
- Reinders, H., & Ditton, H. (2011). Überblick. In H. Reinders, H. Ditton, C. Gräsel & B. Gniewosz (Eds.), *Empirische Bildungsforschung. Strukturen und Methoden* (pp. 45–52). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Röhner, J., & Schütz A. (2016). *Psychologie der Kommunikation*. Wiesbaden: Springer VS.
- Rosenberg, F. v. (2011). *Bildung und Habitustransformation. Empirische Rekonstruktionen und bildungstheoretische Reflexionen*. Bielefeld: Transcript.
- Schäfer, A. (2006). Bildungsforschung: Annäherungen an eine Empirie des Unzugänglichen. In L. Pongratz, M. Wimmer & N. Wolfgang (Eds.), *Bildungsphilosophie und Bildungsforschung* (pp. 87–108). Bielefeld: Janus Presse.
- Schneider, W. L. (2009). *Grundlagen der soziologischen Theorie*. Bd. 2: Garfinkel – RC – Habermas – Luhmann. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Mariño, J. P. (2018). International performance assessment of learning in higher education (iPAL): Research and development. In *Assessment of learning outcomes in higher education* (pp. 193–214). Springer, Cham.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), (pp. 922–932).
- Stevens, R. (2015). Role-play and student engagement: reflections from the classroom. *Teaching in Higher Education*, *20*(5), (pp. 481–492).
- Tillmann, K. J. (2017). Empirische Bildungsforschung in der Kritik – ein Überblick über Themen und Kontroversen. In *Empirische Bildungsforschung* (pp. 5–22). Springer VS, Wiesbaden.
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wesseling, M. (2012). *Der transzendentale Erkenntnisrealismus von Jürgen Habermas*. Marburg: Tectum Verlag.

- Wissenschaftsrat. (2015). *Empfehlungen zum Verhältnis von Hochschulbildung und Arbeitsmarkt. Zweiter Teil der Empfehlungen zur Qualifizierung von Fachkräften vor dem Hintergrund des demographischen Wandels*. WR Publikationen. <https://www.wissenschaftsrat.de/download/archiv/4925-15.pdf>. Accessed: 11 July 2019.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Lautenbach, C., Molerov, D., Toepper, M., & Brückner, S. (2017). *Modeling and Measuring Competencies in Higher Education: Approaches to Challenges in Higher Education Policy and Practice*. Wiesbaden: Springer VS.

Domain-Specific Competencies in Business, Economics and Medicine

4



4.1

Measuring Medical Competence and Entrusting Professional Activities in an Assessment Simulating the First Day of Residency

Prediger, S., Berberat, P. O., Kadmon, M., and Harendza, S.

Abstract

Medical students are supposed to achieve a certain level of entrustment in different professional activities including relevant core and medical competences at the end of their undergraduate studies. To measure relevant facets of competence (FOC) and to evaluate entrusted professional activities (EPA), an assessment, which simulated a first day of residency in a hospital, was carried out twice. In total, 119 students participated once and 11 students participated twice. Linear regression models showed an influence of FOC rating in “Verbal communication with colleagues and supervisors” and “Structure, work planning and priorities” on EPA ratings. Students who participated twice showed significant improvements in these two FOCs and in two EPAs related to these FOCs. The FOC “Responsibility” and the personality characteristics “Extraversion” and “Conscientiousness” positively influenced more difficult entrustment decisions, while “Agreeableness” had a negative influence on their entrustment level. Our assessment format including the FOC and EPA instruments shows good validity and provides important aspects of competence and professionalism, which should be included in teaching and feedback during the final year of undergraduate medical education.

Keywords

Assessment, competences, competence-based assessment, entrustable professional activities (EPA), NEO – Five-Factor Inventory (NEO-FFI), personality characteristics, professionalism, simulation

1 Background

Physicians need to work in a competent and professional way. The learning objectives for undergraduate medical education usually comprise only knowledge, skills, and attitudes, which still leads to ambiguous learning objectives (Guilbert 2002). Competence includes knowledge, skills, and attitudes but additionally requires the ability to use and show those in the clinical environment (Ten Cate 2005). For postgraduate medical education, the CanMEDS framework of competences has been developed integrating seven roles relevant for physicians (Frank and Danoff 2007): (1) Medical Expert (2) Communicator (3) Collaborator (4) Leader (5) Health Advocate (6) Scholar, and (7) Professional. Based on this framework, a National Competence Based Catalogue of Learning Objectives (NKLM) was developed for undergraduate medical studies in Germany, including different facets of competence (FOC), which are operationalized for teaching (Fischer et al. 2015).

Additionally, in the context of competence-based training, the concept of entrustable professional activities (EPA) emerged in medical education (Meyer et al. 2019). Professional activities include different facets of competence, which comprise specific knowledge, skills, and attitudes (Ten Cate 2005). A guide was developed for undergraduate medical education to provide support for EPA-based curriculum development and assessment (Ten Cate et al. 2015).

At the end of their medical studies, graduates are supposed to achieve a certain level of entrustment in different professional activities including relevant core and medical competences. Out of 25 facets of competence, Wijnen-Meijer et al. (2013a) identified ten, for instance, empathy and communication, similarly assessed by physicians from the Netherlands and Germany to be relevant for beginning residents. When we asked physicians teaching at three medical schools with different undergraduate curricula to rank the 25 facets of competence with respect to their importance for beginning residents, eight of the top ten facets of competence from the previous study (Wijnen-Meijer et al. 2013a) were ranked among the top ten again (Fürstenberg et al. 2017). Besides core competences, personality characteristics are relevant factors in the context of medical training (Ferguson et al. 2003) and asso-

ciated with clinical competence (Hojat et al. 2004). Especially conscientiousness shows a relevant influence on long-term success in medical training (Lievens et al. 2009) and additionally predicts vulnerability to stress when combined with a high level of neuroticism and a low level of extraversion (Tyssen et al. 2007).

So far, mini-clinical exercises (mini-CEX) (Norcini et al. 1995), direct observation of procedures (DOPS) (Barton et al. 2012), and objective structured clinical examinations (OSCE) (Khan et al. 2013) have been established as performance-based assessments in undergraduate medical education. While OSCEs, used for summative assessments, are well-structured and standardized, but do not involve real patients, DOPS and mini-CEX appear in non-standardized clinical contexts and are mostly used for formative assessments. With a shift of undergraduate medical education towards competence-based education, a robust and multifaceted assessment system is needed (Holmboe et al. 2010). To test the most relevant facets of competence in advanced medical students, a competence-based assessment format simulating the first day of residency in a hospital had been developed and validated (Wijnen-Meijer et al. 2013b). We redesigned and restructured this assessment (Harendza et al. 2017) as well as extended and validated the instruments for assessing different facets of competence (FOC) and entrustable professional activities (EPA) (Prediger et al. *under review*). The FOC ratings showed good reliabilities (Prediger et al. 2019) and significant correlations between FOCs and EPAs were identified (Fincke et al. *in preparation*).

The aim of this study was to compare the FOC and EPA ratings between two time points of measurement, which differed in content of the patient cases and length of assessment. We also investigated whether personality characteristics contribute explanations to the construct of EPA in correlation with FOCs.

2 Methods

2.1 Setting

To measure facets of competence and evaluate entrusted professional activities, the assessment, which simulated a first day of residency in a hospital, was carried out twice. It took place in July 2018 (t_1) with 70 and in July 2019 (t_2) with 90 advanced medical students from three universities (Hamburg, Oldenburg, and TU Munich) with different undergraduate curricula. A few changes in content and complexity of the assessment were made between t_1 and t_2 , which are described in detail in Figure 1. The adjustments made for t_2 resulted from participants' and raters' feedback.

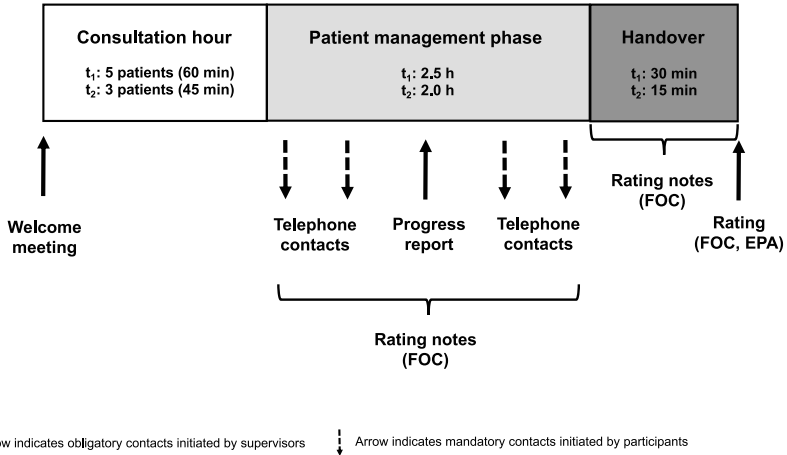


Figure 1 Study Design

The simulation consisted of three phases, which are typical for a resident's daily routine: (1) a consultation hour with five (t₁) or three (t₂) simulated patients; (2) a patient management phase during which the participants requested laboratory tests and interacted with other health care personnel; (3) a handover of patients to a resident (t₁) or another participant (t₂). Each participant in the physician's role had one supervisor who welcomed him/her face-to-face at the beginning of the assessment and met him/her for a short progress meeting half time during the patient management phase. Additionally, participants could call their supervisors on provided cellular phones whenever they wished to discuss any questions with them. Furthermore, supervisors were present as silent observers during the handovers. The supervisors could take rating notes after every contact with their participants. After the handovers, supervisors rated their participants' facets of competence (FOC) on a scoring form with a 5-point scale (1 "insufficient" to 5 "very good", or "no judgement possible") based on their observations and their rating notes. Additionally, the supervisors rated their participants with respect to their entrustment of 12 different entrustable professional activities (EPA), each described in a small case vignette. The EPA scoring form had a 5-point scale (1: no permission to act, 2: permission to act with direct supervision (supervisor present in the room), 3: permission to act with indirect supervision (supervisor not present in the room, but quickly available if needed), 4: permission to act under distant supervision (supervisor not directly available, but a telephone call is possible, i.e. "unsupervised"),

5: permission to provide supervision to junior trainees (Ten Cate et al. 2015)). Additionally, the participants completed the NEO Five-Factor Inventory (NEO-FFI), which is a validated questionnaire to assess the personality domains neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (Costa and McCrae 2008).

2.2 Sample Description

The 119 participants in their final year of a six-year undergraduate medical curriculum who were invited by email and enrolled in the assessment on a first come, first served basis were included in the analysis (t_1 : $n = 41$ and t_2 : $n = 78$). Their mean age was 26.7 ± 3.3 years and 63 % of the 119 students were female. Additionally, 11 medical students (age: 26.3 ± 2.3 years, 54.5 % female), who participated twice in this study, in t_1 at the end of year five and in t_2 at the end of year six, were analysed with respect to their development during their final year, i.e. year six (practice year), where students work full time on hospital wards in internal medicine, surgery, and a specialty of their choice. Students who had neither reached their final year nor participated longitudinally were excluded from the analysis (t_1 : $n = 29$ and t_2 : $n = 1$).

2.3 Data Analysis

For statistical analysis, we conducted independent-samples t-tests for all FOCs and EPAs using SPSS Statistics 25 comparing the results in t_1 with the results in t_2 for the 119 participants. To explore the development of the 11 medical students between t_1 and t_2 , we conducted the Wilcoxon-Test as a non-parametric test due to the small sample size. We also computed *Cohen's d* for all mean comparisons to explore effect sizes. For graphical representation of medians and scattering in EPA ratings for participants, who participated only once in our assessment, we provide boxplots, which show, which EPAs were assessed with higher or lower levels of entrustment than others. In accordance with this, we computed a variable based on EPAs, which were entrusted on average with a lower level (*EPA_low*) and a variable based on EPAs, which were entrusted on a higher level (*EPA_high*). To explore potential differences in explanation of these variables, we modelled *EPA_low* and *EPA_high* as two multiple linear regressions with single FOCs and personality characteristics as regressors. Additionally, we controlled for sex and time points of measurement (t_1 or t_2).

3 Results

With respect to FOCs, one significant difference between t_1 and t_2 could be detected (Table 1). Participants scored lower in “Scientifically and empirically grounded method of working” in t_2 . This difference, -0.75 , BCa 95 % CI [.101, 1.400], was significant ($t(45.6) = 2.328, p = .024$) with a medium-sized effect ($d = .622$).

Table 1 Comparison of FOC ratings between the student cohorts at t_1 and t_2

		<i>M</i> ± <i>SD</i>	<i>N</i>	<i>p</i>
Responsibility	t_1	3.63 ± .97	35	.924
	t_2	3.65 ± 1.10	77	
Teamwork and collegiality	t_1	3.71 ± .76	34	.719
	t_2	3.78 ± 1.00	67	
Knowing and maintaining own personal bounds and possibilities	t_1	3.37 ± 1.09	41	.678
	t_2	3.46 ± 1.22	76	
Structure, work planning and priorities	t_1	3.32 ± 1.21	41	.171
	t_2	3.62 ± 1.07	78	
Coping with mistakes	t_1	3.61 ± .93	33	.493
	t_2	3.76 ± 1.09	51	
Scientifically and empirically grounded method of working	t_1	3.45 ± 1.00	33	.024
	t_2	2.70 ± 1.42	27	
Verbal communication with colleagues and supervisors	t_1	3.71 ± 1.06	41	.713
	t_2	3.63 ± 1.14	78	

Furthermore, supervisors evaluated participants in t_2 in the EPA “Medication error” (EPA 11) on average on a higher level of entrustment (Table 2). This difference, 0.45 , BCa 95 % CI [-.798, -.101], was significant ($t(113) = -2.556, p = .012$) with a medium effect size ($d = .500$).

We found two significant differences for the 11 medical students, who participated twice (Table 3), with high effect sizes ($d = 2.980$ and $d = 2.440$). These participants improved in their facets of competence “Structure, work planning and priorities” ($\tau = -2.75, p = .006$) and “Verbal communication with colleagues and supervisors” ($\tau = -2.57, p = .010$). Moreover, supervisors assessed them longitudinally on average on a higher level of entrustment in two EPAs: “Emergency assistance in a case with acute cardiac failure” ($\tau = -2.50, p = .013$) and “Solving a management problem” ($\tau = -2.08, p = .037$) (Table 4). Both differences have high effect sizes, $d = 2.286$ and $d = 1.612$, respectively. In accordance with the distribution of EPA ratings (Figure 2) we computed *EPA_low* from the means of

the EPAs “Emergency assistance in a case with acute cardiac failure” (EPA 1), “Handling of a critically ill patient” (EPA 8), and “Acting to patient’s will” (EPA 12), as those were entrusted on average with a lower level (1–2), which means participants would get no permission to act or only the permission to act with direct supervision. *EPA_low* is linearly predicted by the facets of competence “Structure, work planning and priorities” and “Verbal communication with colleagues and supervisors” as well as by the personality characteristics “Extraversion”, “Agreeableness”, and “Conscientiousness” (Table 5a). Those five variables significantly predict lower entrustment of EPAs ($F(5, 113) = 18.59, p < .001$). The R^2 for the overall model is .45 (adjusted $R^2 = .43$), indicating a high goodness-of-fit.

Table 2 Comparison of EPA ratings between the student cohorts at t_1 and t_2

		<i>M</i> ± <i>SD</i>	<i>N</i>	<i>p</i>
EPA 1: Emergency assistance in a case with acute cardiac failure	t_1	1.66 ± .62	41	.770
	t_2	1.69 ± .59	78	
EPA 2: Handling a patient’s complaint	t_1	2.10 ± .86	41	.093
	t_2	2.40 ± 1.01	78	
EPA 3: Pre-operative information and consent	t_1	3.12 ± .68	41	.413
	t_2	3.24 ± .91	78	
EPA 4: Breaking bad news	t_1	2.05 ± .71	41	.748
	t_2	2.00 ± .82	78	
EPA 5: Clinical reasoning under time pressure	t_1	2.24 ± .73	41	.061
	t_2	1.97 ± .77	78	
EPA 6: Solving a management problem	t_1	3.44 ± .67	41	.197
	t_2	3.64 ± .84	77	
EPA 7: Suspicion of self-induced disease	t_1	2.61 ± .83	41	.102
	t_2	2.30 ± 1.03	78	
EPA 8: Handling of a critically ill patient	t_1	1.71 ± .64	41	.793
	t_2	1.67 ± .69	78	
EPA 9: Interaction with a consultant	t_1	3.17 ± .83	41	.766
	t_2	3.22 ± .82	78	
EPA 10: Presentation of an oncology patient in a tumor board meeting	t_1	3.02 ± .88	41	.096
	t_2	2.72 ± .98	78	
EPA 11: Medication error	t_1	2.78 ± .88	41	.012
	t_2	3.23 ± .91	74	
EPA 12: Acting to patient’s will	t_1	1.63 ± .70	41	.078
	t_2	1.39 ± .74	78	

Table 3 Comparison of FOC ratings between t_1 and t_2 of medical students with longitudinal participation

		$M \pm SD$	N	p
Responsibility	t_1	3.57 ± .79	7	.180
	t_2	4.14 ± .69		
Teamwork and collegiality	t_1	3.44 ± .53	9	.084
	t_2	4.11 ± .78		
Knowing and maintaining own personal bounds and possibilities	t_1	3.36 ± 1.03	11	.465
	t_2	3.64 ± 1.03		
Structure, work planning and priorities	t_1	2.73 ± .79	11	.006
	t_2	3.91 ± .83		
Coping with mistakes	t_1	3.44 ± .53	9	.157
	t_2	3.89 ± .60		
Scientifically and empirically grounded method of working	t_1	3.00 ± 1.41	2	.180
	t_2	4.50 ± .71		
Verbal communication with colleagues and supervisors	t_1	3.09 ± .83	11	.010
	t_2	4.27 ± .79		

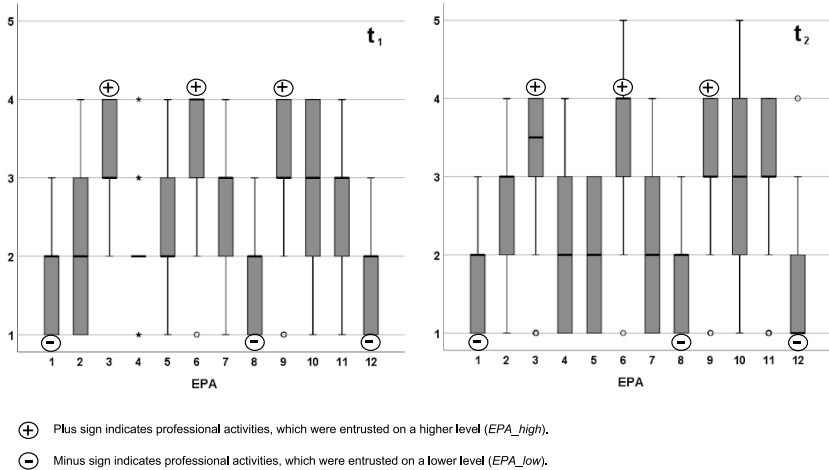


Figure 2 Distribution of EPA ratings

We computed *EPA_high* from the means of the EPAs “Pre-operative information and consent” (EPA 3), “Solving a management problem” (EPA 6), and “Interaction with a consultant” (EPA 9), as those were entrusted on a higher level (3–4), which means participants would get permission to act with indirect or under distant supervision. *EPA_high* is also linearly predicted by the two facets of competence “Structure, work planning and priorities” and “Verbal communication with colleagues and supervisors”. Additionally, the competence facet “Responsibility” explains a part of the variance (Table 5b). Those three variables significantly predict higher entrustment of EPAs ($F(3, 108) = 37.49, p < .001$). No personality characteristics explained part of the variance for *EPA_high*. The R^2 for the overall model is .51 (adjusted $R^2 = .50$), indicating a high goodness-of-fit. Sex and time points of measurement (t_1 or t_2) had no significant effects on *EPA_high* or *EPA_low*.

Table 4 Comparison of EPA ratings between t_1 and t_2 of medical students with longitudinal participation

		<i>M</i> ± <i>SD</i>	<i>N</i>	<i>p</i>
EPA 1: Emergency assistance in a case with acute cardiac failure	t_1	1.45 ± .69	11	.013
	t_2	2.27 ± .47		
EPA 2: Handling a patient’s complaint	t_1	2.18 ± .87	11	.102
	t_2	2.82 ± .75		
EPA 3: Pre-operative information and consent	t_1	2,64 ± .81	11	.414
	t_2	2.82 ± .87		
EPA 4: Breaking bad news	t_1	2.18 ± .98	11	.618
	t_2	1.91 ± .83		
EPA 5: Clinical reasoning under time pressure	t_1	2.18 ± .87	11	.276
	t_2	2.45 ± .52		
EPA 6: Solving a management problem	t_1	3.45 ± .69	11	.037
	t_2	4.27 ± .79		
EPA 7: Suspicion of self-induced disease	t_1	2,45 ± .82	11	.792
	t_2	2.55 ± .82		
EPA 8: Handling of a critically ill patient	t_1	1.64 ± .67	11	.180
	t_2	1.91 ± .54		
EPA 9: Interaction with a consultant	t_1	3.09 ± .70	11	.705
	t_2	3.18 ± .75		
EPA 10: Presentation of an oncology patient in a tumor board meeting	t_1	2.82 ± .87	11	1.000
	t_2	2.82 ± .87		
EPA 11: Medication error	t_1	2.64 ± .67	11	.557
	t_2	2.82 ± .98		
EPA 12: Acting to patient’s will	t_1	1.73 ± .79	11	.160
	t_2	1.27 ± .65		

Table 5a Linear regression Model of *EPA_low*

	<i>b</i>	<i>SD</i>	β	<i>p</i>
(Constant)	.198	.310		.524
Structure, work planning and priorities	.158	.049	.349	.002
Verbal communication with colleagues and supervisors	.143	.050	.312	.005
Extraversion	.015	.007	.175	.032
Agreeableness	-.016	.007	-.173	.032
Conscientiousness	.012	.006	.150	.038

Table 5b Linear regression Model of *EPA_high*

	<i>b</i>	<i>SD</i>	β	<i>p</i>
(Constant)	1.619	.169		.000
Responsibility	.152	.067	.243	.027
Structure, work planning and priorities	.157	.073	.261	.034
Verbal communication with colleagues and supervisors	.164	.070	.275	.021

4 Discussion and Conclusion

Comparing the FOC and EPA ratings between two different cohorts of final year medical students at two different time points with certain adjustments in the overall assessment format, we discovered only two significant differences. The FOC “Scientifically and empirically grounded method of working” was rated lower in t_2 and the EPA “Medication error” was entrusted on a higher level in t_2 . These differences might be due to introducing anchors in the FOC scoring form in t_2 (Prediger et al. *under review*). None of the other ratings was influenced by changes made between the two time points of measurement, which indicates good validity of the assessment format.

Medical students, who participated twice in the assessment, showed higher ratings for all FOCs with significant increases in “Verbal communication with colleagues and supervisors” and “Structure, work planning and priorities”. At t_1 , we discovered that students had received the lowest FOC rating for “Structure, work planning and priorities” (Prediger et al. 2019) and felt the highest strain during the patient management phase of the assessment where this FOC plays an important role (Fürstenberg et al. 2018). Higher FOC ratings at t_2 could be due to participants’ experience of responsibility in our assessment, which many reported

to have “really felt” for the first time and which might have stimulated them to set a specific learning focus on those two core FOCs for medical practice during their practice year. Additionally, they had received individual feedback on their FOC ratings after t_1 (Harendza et al. 2017), which showed a need for improvement, too. The improvement in entrustment decisions could be due to students’ acquisition or refinement of competences during their practice year, which address different facets of competence included in our EPAs as learning objectives. To enhance core and medical competences during the practice year, where medical students participate fulltime in clinical work on hospital wards, different learning formats have been implemented including, for example, video based on-ward supervision and feedback (Groener et al. 2015) or portfolios (Zundel et al. 2015). With a multiple choice test, conducted at the beginning and at the end of the practice year, an increase in medical knowledge could also be shown in medical students at the end of their final year (Raupach et al. 2013).

EPAs for which participants received lower entrustment levels (*EPA_low*) and EPAs for which they received higher entrustment levels (*EPA_high*) are both predicted linearly by the same FOCs, namely “Structure, work planning and priorities” and “Verbal communication with colleagues and supervisors”. Interestingly, these are the two FOCs where medical students who participated twice showed a significant increase. Apparently, supervisors’ observation of the extent of these two core competences plays an important role for the level of entrustment of professional activities in general, even though usually every EPA is based on a specific set of competences (Mulder et al. 2010). These two core competences were shown to be of importance for most of the EPAs used in our study (Fincke et al. *in preparation*). For *EPA_high*, the competence facet “Responsibility” explains a part of the variance. “Responsibility” received the highest scores in our ranking study (Fürstenberg et al. 2017) and was also among the five most important non-cognitive goals which should be achieved during undergraduate medical observation (Mann et al. 2005). Interestingly, only *EPA_low* is additionally predicted by personality characteristics, i.e. “Extraversion” and “Conscientiousness”, which predict higher entrustment levels while “Agreeableness” predicts lower entrustment levels for *EPA_low*. Another study showed that a conscientiousness index in medical students highly correlates with the students’ professionalism (McLachlan et al. 2009), which might also explain its higher correlation with difficult EPAs in our study. Extraversion in medical students has been found to predict their empathic communication (Schreckenbach et al. 2018) and seems to be an additional factor to the FOC “Verbal communication with colleagues and supervisors” to predict higher ratings for *EPA_low*. “Agreeableness”, which has been found to be associated with empathy (O’Tuathaigh et al. 2019), can be a hindrance when dealing

with emergency EPAs, which could explain its negative prediction of *EPA_low*. High scores for “Agreeableness” have also been shown to be associated with a delay in undergraduate medical studies (Walldorf and Fischer 2018). Personality characteristics were presumed to be invariable from the age of 30 (McCrae and Costa 1994). However, more recent research assumes that they may change even in adulthood, i.e. to increase self-confidence, self-control, and emotional stability (Roberts and Mroczek 2008). Hence, personality characteristics should also become a topic of teaching in undergraduate medical education, for instance, with exercises in self-reflection and awareness.

One strength of our study is the high overall number of participants and the stable assessment of FOCs and EPAs at t_1 and t_2 suggesting good validity of this assessment format. A limitation is the small number of students who participated twice. However, their results show first tendencies of competence development during the final year of undergraduate medical education which are encouraging to use this new format as teaching and assessment tool. Further investigations with a larger number of longitudinal participants are necessary to underscore our initial findings. Additionally, specific teaching formats could be developed for the final year of undergraduate medical education to focus on the enhancement of the top ten facets of competence for beginning residents.

Our newly developed instruments to measure facets of competence and entrustable professional activities show stable ratings in two final year student cohorts despite changes in duration of the assessment and in the number of patient cases. When students were assessed twice, significant improvement in the assessment of specific FOCs and EPAs could be observed. Developing the FOCs further during the final year and the observed influence of personal characteristics on entrustment decisions for more difficult EPAs should become a nucleus for teaching and feedback in undergraduate medical education.

5 Acknowledgements

This study was part of the project ÄKHOM, which was funded by the German Ministry of Education and Research (BMBF), reference number: 01PK1501A/B/C. We would like to thank all physicians and medical students from the Universitätsklinikum Hamburg-Eppendorf, from the Carl von Ossietzky University of Oldenburg, and from the Technical University of Munich, who participated in this study.

References

- Barton, J. R., Corbett, S., & an der Vleuten, C. P. (2012). The validity and reliability of a direct observation of procedural skills assessment tool: Assessing colonoscopic skills of senior endoscopists. *Gastrointestinal Endoscopy* 75, pp. 591–597.
- Costa, P. T., & McCrae, R. R. (2008). The Revised Neo Personality Inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment* 2, pp. 179–198.
- Ferguson, E., James, D., O’Hehir, F., Sanders, A., & McManus, I. C. (2003). Pilot study of the roles of personality, references, and personal statements in relation to performance over the five years of a medical degree. *BMJ* 326(7386), pp. 429–432.
- Fincke, F., Prediger, S., Schick, K., Fürstenberg, S., Spsychala, N., Berberat, P. O., Harendza, S., & Kadmon, M. (in preparation). Entrustable professional activities and facets of competence in a simulated workplace-based assessment for advanced medical students. *Medical Teacher*.
- Fischer, M. R., Bauer, D., & Mohn, K. (2015). Finally finished! National competence based catalogues of learning objectives for undergraduate medical education (NKLM) and dental education (NKLZ) ready for trial. *GMS Zeitschrift für Medizinische Ausbildung* 32, Doc35.
- Frank, J. R., & Danoff, D. (2007). The CanMEDS initiative: implementing an outcomebased framework of physician competencies. *Medical Teacher* 29, pp. 642–647.
- Fürstenberg, S., Prediger, S., Kadmon, M., Berberat, P. O., & Harendza, S. (2018). Perceived strain of undergraduate medical students during a simulated first day of residency. *BMC Medical Education* 18(1), pp. 322.
- Fürstenberg, S., Schick, K., Deppermann, J., Prediger, S., Berberat, P. O., Kadmon, M., & Harendza, S. (2017). Competencies for first year residents. physicians’ views from medical schools with different undergraduate curricula. *BMC Medical Education* 17(1), pp. 154.
- Groener, J. B., Bugaj, T. J., Scarpone, R., Koechel, A., Stiepak, J., Branchereau, S., Krautter, M., Herzog, W., & Nikendei, C. (2015). Video-based on-ward supervision for final year medical students. *BMC Medical Education* 15, pp. 163.
- Guilbert, J. J. (2002). The ambiguous and bewitching power of knowledge, skills and attitudes leads to confusing statements of learning objectives. *Education for Health* 15(3), pp. 362–369.
- Harendza, S., Berberat, P. O., & Kadmon, M. (2017). Assessing competences in medical students with a newly designed 360-degree examination of a simulated first day of residency: a feasibility study. *Journal of Community Medicine and Health Education* 7, pp. 4.
- Hojat, M., Callahan, C. A., & Gonnella, J. S. (2004). Students’ personality and ratings of clinical competence in medical school clerkships: a longitudinal study. *Psychology, Health & Medicine* 9(2), pp. 247–252.
- Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher* 32(8), pp. 676–682.
- Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013). The objective structured clinical examination (OSCE): AMEE guide no. 81. Part II: organisation & administration. *Medical Teacher* 35(9), pp. 1447–1463.
- Lievens, P., Ones, D. S., & Dilchert, S. (2009). Personality scale validities increase throughout medical school. *Journal of Applied Psychology* 94(6), pp. 1514–1535.

- Mann, K. V., Ruedy, J., Millar, N., & Andreou, P. (2005). Achievement of non-cognitive goals of undergraduate medical education: perceptions of medical students, residents, faculty and other health professionals. *Medical Education* 39(1), pp. 40–48.
- McLachlan, J. C., Finn, G., & Macnaughton, J. (2009). The conscientiousness index: a novel tool to explore students' professionalism. *Academic Medicine* 84(5), pp. 559–565.
- McCrae, R. R., & Costa, P. T. (1994). The stability of personality: Observation and evaluations. *Current Directions in Psychological Science* 3, pp. 173–175.
- Meyer, E. G., Chen, H. C., Uijtdehaage, S., Durning, S. J., & Maggio, L. A. (2019). Scoping Review of Entrustable Professional Activities in Undergraduate Medical Education. *Academic Medicine* 94(7), pp. 1040–1049.
- Mulder, H., Ten Cate, O., Daalder, R., & Berkvens, J. (2010). Building a competency-based workplace curriculum around entrustable professional activities: The case of physician assistant training. *Medical Teacher* 32(10), pp. 453–459.
- Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine* 123, pp. 795–799.
- O'Tuathaigh, C. M. P., Nadhirah Idris, A., Duggan, E., Costa, P., & Costa, M. J. (2019). Medical students' empathy and attitudes towards professionalism: Relationship with personality, specialty preference and medical programme. *PLoS One* 14(5), pp. 215–675.
- Prediger, S., Schick, K., Fincke, F., Fürstenberg, S., Oubaid, V., Kadmon, M., Berberat, P. O., & Harendza, S. (under review). Validation of a competence-based assessment of medical students' performance in the physician's role. *BMC Medical Education*.
- Prediger, S., Fürstenberg, S., Berberat, P. O., Kadmon, M., & Harendza, S. (2019). Interprofessional assessment of medical students' competences with an instrument suitable for physicians and nurses. *BMC Medical Education* 19(1), p. 46.
- Raupach, T., Vogel, D., Schiekirka, S., Keijsers, C., Ten Cate, O., & Harendza, S. (2013). Increase in medical knowledge during the final year of undergraduate medical education in Germany. *GMS Zeitschrift für Medizinische Ausbildung* 30(3), Doc33.
- Roberts B. W., & Mroczek, D. (2008). Personality Trait Change in Adulthood. *Current Directions in Psychological Science* 17(1), pp. 31–35.
- Schreckenbach, T., Ochsendorf, F., Sterz, J., Rüsseler, M., Bechstein, W. O., Bender, B., & Bechtoldt, M. N. (2018). Emotion recognition and extraversion of medical students interact to predict their empathic communication perceived by simulated patients. *BMC Medical Education* 18(1), p. 237.
- Ten Cate, O. (2005). Entrustability of professional activities and competency-based training. *Medical Education* 39(12), pp. 1176–1177.
- Ten Cate, O., Chen, H. C., Hoff, R. G., Peters, H., Bok, H., & van der Schaaf, M. (2015). Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. *Medical Teacher* 37(11), pp. 983–1002.
- Ten Cate, T. J. O., Snell, L., & Carraccio, C. (2010). Medical competence: the interplay between individual ability and the health care environment. *Medical Teacher* 32(8), pp. 669–675.
- Tyssen, R., Dolatowski, F. C., Røvik, J. O., Thorkildsen, R. F., Ekeberg, O., Hem, E., Gude, T., Gronvold, N. T., & Vaglum, P. (2007). Personality traits and types predict medical school stress: a 6-year longitudinal and nationwide study. *Medical Education* 41(8), pp. 781–787.

- Walldorf, J., & Fischer, M. R. (2018). Risk factors for a delay in medical education: Results of an online survey among four German medical schools. *Medical Teacher* 40(1), pp. 86–90.
- Wijnen-Meijer, M., van der Schaaf, M., Nillesen, K., Harendza, S., & Ten Cate, O. (2013a). Essential facets of competence that enable trust in medical graduates: a ranking study among physician educators in two countries. *Perspectives on Medical Education* 2(5–6), pp. 290–297.
- Wijnen-Meijer, M., Van der Schaaf, M., Booij, E., Harendza, S., Boscardin, C., Van Wijngaarden, J., & Ten Cate, T. J. (2013b). An argument-based approach to the validation of UHTRUST: Can we measure how recent graduates can be trusted with unfamiliar tasks? *Advances in Health Science Education. Theory and Practice* 18, pp. 1009–1027.
- Zundel, S., Blumenstock, G., Zipfel, S., Herrmann-Werner, A., & Holderried, F. (2015). Portfolios enhance clinical activity in surgical clerks. *Journal of Surgical Education* 72(5), pp. 927–935.



4.2

Impact of Affective-Motivational Dispositions on Competence in Sustainability Management¹

Michaelis, C., Aichele, C., Hartig, J., Seeber, S., Dierkes, S., Schumann, M., Jan Moritz, A., Siepelmeyer, D., and Repp, A.

Abstract

Only limited evidence exists to explain competence in sustainability management. In sustainability research, especially value-related constructs are emphasized to predict sustainable behaviour. In addition, theoretical-conceptual considerations toward competence and some empirical analyses of competence measurement highlight the potential of affective-motivational dispositions to explain the performance in competence tests. Therefore, this paper measures the influence of affective-motivational dispositions toward the performance of competence in sustainability management. To measure the performance of competence in sustainability management, a simulation-based assessment is used. Based on existing findings of previous research scales toward the declarative knowledge about sustainable development, sustainability management and business administration are used as control variables. The analyses based on responses from 872 students and are done by structural equation models. Besides the declarative knowledge about sustainable development, an aversion to sustainability has a significant influence on the performance of competence in sustainability management. The motivation to act sustainably as well as interest-based constructs show no significant results.

¹ This project is part of the funding initiative “Modeling and Measuring Competencies in Higher Education (KoKoHs)”, funded by the Federal Ministry of Education and Research under grant no. 01PK15010A. Responsibility for the content of this publication lies with the author.

Keywords

Competence model in sustainability management, competence structure, competence diagnostics, affective-motivational dispositions, attitudes toward sustainability, motivation to act sustainable, simulation-based assessment, corporate social responsibility

1 Introduction

In recent years, the debate about the importance of higher education for a sustainable development has intensified (Lozano et al. 2015). The focus of this perspective is the promotion of skills that can satisfy the needs of the present generation without risking the fundamentals of life for the next generation (in line with the main goals of sustainable development: United Nations, n.d.). Although numerous competence models supporting a sustainable development are discussed for higher education, their focus is on comprehensive generic action, shaping, problem-solving or key competences (for an overview, see Hesselbart and Schaltegger 2014, Lozano et al. 2017, Wiek et al. 2011). Theoretically assumed dimensions in these models are often difficult to operationalize and structural relationships are inadequately defined.² Therefore, most approaches are closed to an empirical verification according to assumptions of competence diagnostics (Hartig et al. 2008). In addition, domain specific facets are largely neglected in these models. This paper focuses on a domain-specific model for sustainability management and is not referring to comprehensive generic competence constructs of higher education for sustainable development.

Competences in sustainability management are considered as a central domain for promoting a sustainable development due to potential participation of students after graduation in strategic decisions in companies (Seeber et al. 2019). Sustainability management aims to promote a sustainable development by reconciling it with assumptions of business administration. However, this is a challenge due to numerous conflicts between economic, environmental and social entrepreneurial goals according to expectations of different stakeholders (especially investors, customers, suppliers, employees, state, etc.). Therefore, the focus of sustainability management is on the development of proposed solutions for corresponding con-

2 See, for example, Gräsel et al.'s (2012) critics toward the shaping competence, which is widespread in the education for sustainable development.

flicting goals in an entrepreneurial context. With regard to the definition of competences in sustainability management, we follow Seeber et al. (2019):

„We define competence in sustainability management as the complex ability to identify and consider the stakeholders’ partly joining, partly conflicting economic, environmental and social goals in the target system of a company. It means in particular to be able to take into account the short-, medium- and long-term interactions of the stakeholders’ different goals and their consequences for the company as well as for the company’s surrounding. Therefore, sustainability management means to manage a company in a way that it exists in a long term with a positive contribution of the company to the sustainable development of society and natural conditions.“

Sustainability management is still an emerging field of research and teaching of business administration. The development and validation of competence models are still in their infancy. Empirical evidence on competence of sustainability management has hitherto been available only on the structural relationships of individual knowledge dimensions. Seeber et al. (2019) show that the declarative knowledge about sustainability, incorporated into a general societal perspective, has the strongest correlation with the ability to generate strategies and justifications for specific options in terms of sustainability management. In addition, declarative knowledge of business administration and declarative knowledge of sustainability management also have positive correlations with the ability to generate strategies and justifications for specific options in terms of sustainability management. However, these coefficients are weaker.

In particular, findings from environmental and sustainability awareness research repeatedly show the importance of values and attitudes on environmental or sustainability decisions (Bamberg and Möser 2007, Leiserowitz et al. 2006, Kaiser et al. 1999, Shepherd et al. 2009). In addition, research in competence diagnostics highlights the potential of affective-motivational dispositions (like attitudes, motivation and interest) in domain-specific decision-making situations (Blömeke et al. 2015). Besides, a positive influence of sustainability-related attitudes on the competence to act sustainably at the workplace (Seeber and Michaelis 2014) and its development (Michaelis 2017) have been found for trainees of the business domain. So far, there is no evidence for the influence of affective-motivational dispositions on competence in sustainability management. We address this research gap in this article and combine this with Seeber et al.’s findings (2019). We focus on the following research question: Can affective-motivational dispositions predict competence in sustainability management?

2 Theoretical-Conceptual Background and Hypotheses

Against the background of theoretical conceptions of competence, the performance in domain-specific requirement situations is assumed to be based primarily on underlying cognitive as well as affective-motivational dispositions (Blömeke et al. 2015). Analyses of domain-specific competences in economics and business administration have so far focused mainly on subject-related cognitive facets of competence (knowledge aspects, e.g., Happ et al. 2016 a, b; analyses toward the meaning of mental processes, e.g., Brückner and Pellegrino 2017). In some studies, motivational aspects are taken into account for the explanation of variances in competence or knowledge development (e.g., Biewen et al. 2018). However, affective dispositions are widely not considered in these analyses. One reason for this is the high importance of economic-rational reasoning in management situations. All the more important is the knowledge of domain-specific facts and procedures. Motivational dispositions are in particular considered as an indicator for the development of any competences (Biewen et al. 2018). Sustainability management is often characterized by a high level of interdisciplinary complexity, although business administration procedures are also considered to be decisive (Seeber et al. 2019). However, in sustainability management, decision-making takes place under a high level of uncertainty. Individual affective-motivational dispositions (especially sustainability-related attitudes, interests toward sustainability topics, and motivation to act sustainably) become important for decision-making processes in sustainability management (Michaelis 2017).

Attitudes are latent dispositions that serve the purpose of individual evaluation of objects, persons, or events (Ajzen 1989). Attitude measurement has a long tradition in sustainability and especially environmental awareness research and is often associated with a behavioural control function. Considering theoretical assumptions of behavioural models (e.g., Ajzen 1991), it is assumed that attitudes are one of the main indicators for predicting a specific sustainable behaviour. For example, in meta study analyses a weak to medium correlation is observed for predicting environmentally intentions by pro-environmental attitudes (Bamberg and Möser 2007). High correlations should not be expected, as a wide repertoire of values can influence the decision-making process in sustainability problems (Shepherd et al. 2009). This variety of underlying values is rarely considered in existing analyses. Moreover, discrepancies between sustainability-related attitudes and sustainable behaviour are observed (e.g., Moser and Kleinhüttelkotten 2018). Behavioural barriers can be assumed as an explanation for the finding that sustainability-related attitudes do not match perfectly with sustainable behaviour. Sustainability aspects might be given a lower priority than other crucial values,

behavioural control aspects might be lacking (such as time, money, knowledge or efficacy) or new behaviours might be stand against established routines. Perceived social norms or economic-political context factors such as laws, lacking infrastructure or technological progress (Leiserowitz et al. 2006) are worth mentioning, as these aspects can also limit the realisation of sustainability-related attitudes to concrete sustainable behaviour. Also, in the context of sustainability management, sustainability-related attitudes do not necessarily lead to higher performance in competence of sustainability management (Michaelis 2017). Thus, in an entrepreneurial context, not every investment in environmental or social aspects is compatible with corporate goals and depends on synergy effects to an economic benefit (Schaltegger and Synnestvedt 2002). Corresponding rationality assumptions could also limit the impact of sustainability-relevant attitudes in management decisions.

Motivational dispositions, as well as interest, have also to be regarded as dispositions which might explain variances in competence of sustainability management. In comparison to the influence of values and attitudes on sustainability decisions, however, the influence of motivation and interest on sustainability-related decision-making has been less systematically investigated. One reason may lie in the breadth of potential theoretical foundations in motivational research (Heckhausen and Heckhausen 2018). In competence diagnostics assumptions of the self-determination theory of Ryan and Deci (2004) are common. It is assumed that individuals strive for competence, autonomy, and social inclusion. These needs are important for motivational factors. People are considered to be motivated if they pursue a specific goal. Closely linked to motivation is the construct of interest that is described as a meaningful relationship of a person to an object. This may also relate to specific contents, activities, concrete objects or abstract ideas (Lewalter et al. 2001).

Taking into account the theoretical and conceptual considerations, the following hypotheses are examined.

H1: Positive *attitudes towards sustainability* have a positive effect on competence in sustainability management when controlling for declarative knowledge.

H2: The *motivation to act sustainably* has a positive effect on competence in sustainability management when controlling for declarative knowledge.

H3: *The interest in sustainability topics* has a positive effect on competence in sustainability management when controlling for declarative knowledge.

3 Methodological Approaches

3.1 Sample

The sample is mainly identical with the one used by Seeber et al. (2019) but comprises data from additional participants. We tested 872 students (375 female, 317 male, 180 missing) in 16 higher education institutions in Germany. To ensure a representative sample, the majority of students were tested directly in lectures and tutorial sessions. Participation was voluntary in accordance with the legal requirements. To enhance test motivation, all participants received a voucher for cinema or online shopping. The test was conducted by tablet computers, testing time was 90 minutes.

The students were mainly enrolled in economics and business administration programs (59%), predominantly in the higher semester of a Bachelor or a Master course of studies. Students were between 18 and 38 years old (*median* = 23, *SD* = 3.05), most of them studying a bachelor's program (483 bachelor, 206 master, 183 missing).

3.2 Instruments

3.2.1 Performance of Competence in Sustainability Management and Declarative Knowledge Tests

To measure the performance of competence in sustainability management a simulation-based assessment is used which was developed by Seeber et al. (2019). This instrument measures the ability to generate strategies and justifications for specific options in terms of sustainability management. In this instrument, the students have to empathize with authentic management situations and have to make and justify decisions for different problems and challenges of a bicycle producer in the sense of sustainability management. All situations are initiated by a video vignette and associated items are supported by different authentic workplace materials (e-mails, spreadsheets, presentations, etc.). From a psychological perspective, this includes particularly the measurement of schematic [why] and strategic [when, where, how] knowledge (based on Shavelson et al. 2005), why this instrument is called SSKSM (schematic and strategic knowledge in sustainability management) by Seeber et al. (2019). However, due to the situational anchoring, this instrument could also indicate the performance of competence in sustainability management. The SSKSM instrument comprises a total of 13 different situations with 73 items.

In the analysis of Seeber et al. (2019) three additional knowledge tests were used. The declarative knowledge about sustainable development from a societal perspective (KSD) instrument contains a total of 53 single and multiple-choice items. The declarative knowledge in business administration (KBA) instrument contains 80 single and multiple-choice items. The declarative knowledge in sustainability management (KSM) contains 51 single and multiple-choice items (for more details regarding the knowledge instruments, see Seeber et al. 2019).

3.2.2 Questionnaire Scales

To measure sustainability-related attitudes, interests, and motivation, we used different questionnaire scales. The scales have a 4 point Likert scale and range from 1 (not true at all) to 4 (fully true). Two scales are used to measure sustainability-related attitudes. In the sense of sustainability management, one scale (Table 1) examines the extent to which students agree that companies have to take responsibility for sustainability (*ia.comp.res.sust*). This is an adaptation of Seeber and Michaelis (2014) and Michaelis (2017). A second scale (Table 2) measures an aversion toward sustainability (*aversion*). Due to the high societal relevance of sustainability issues, it was deliberately aimed to measure a negative disposition to counteract socially desirable responses. Due to the negative formulation, a negative effect must be expected regarding H1. The scale for measuring the motivation to act sustainably (*mot.sust*, Table 5) based on a scale by Michaelis (2017). However, the items were slightly adapted linguistically for this study.

CFAs were performed to examine one-dimensional structures for those questionnaire scales that can be modelled in overidentified models with a minimum of four items. Interest in sustainability topics didn't show satisfactory fit in one-dimensional CFA ($\chi^2 = 193.02$; $df = 5$; $p < .001$; RMSEA = .25; CFI = .94; TLI = .88). An explanatory factor analysis revealed that a two factor solution would describe the data better (Tab. A.1). Based on the loadings in the explanatory factor analysis, we divided the scale into two dimensions ($\chi^2 = 21.622$; $df = 4$; $p < .001$; RMSEA = .086; CFI = .99; TLI = .99) that address 1) *social* (*int.chal.dev.count.*, Table 3) and 2) *ecological* sustainability topics (*int.eco.sust.*, Table 4).

Table 1 Individual attitude that companies have to take responsibility for sustainability (ia.comp.res.sust)

Item	Discrimination	M	SD	Difficulty
Sustainability management in companies (this means the alignment of all activities to the targets of economy-ecology-social affairs as well as their interrelations) is a delightful task.	0.49	3.33	0.72	0.78
Companies should produce their products sustainable, even if prices increase.	0.55	2.98	0.76	0.66
Companies should promote sustainability more with their customers.	0.57	3.19	0.75	0.73

Note. mean score = 3.16; standard deviation scale = 0.59; Cronbach's Alpha α = 0.715; measurement was in German language.

Table 2 Aversion to sustainability (aversion)

Item	Discrimination	M	SD	Difficulty
Sustainability management in companies is, in my opinion, a pure luxury problem.	0.61	1.64	0.83	0.21
If there are more regulations for nature conservation, you will not be able to do anything at all soon.	0.59	1.84	0.87	0.28
Fairtrade does not really help to improve living conditions in developing countries.	0.37	2.25	0.93	0.42
The ongoing discussion on sustainability issues is becoming annoying.	0.61	1.84	0.89	0.28
To my mind, problems in developing countries are greatly exaggerated by stakeholders.	0.58	1.69	0.82	0.23

Note. mean score = 1.86; standard deviation scale = 0.63; Cronbach's Alpha α = 0.776; CFI = 1; TLI = 1; RMSEA < .001; measurement was in German language.

Table 3 Interest in topics about challenges in developing countries (int.chal.dev.count.)

Item	Discrimination	M	SD	Difficulty
How much are you interested in...				
... economic problems in developing countries.	0.64	2.81	0.77	0.60
... ecological problems in developing countries (e.g. effects of monocultures).	0.60	2.61	0.79	0.54
... social problems in developing countries (e.g. working conditions).	0.58	2.93	0.79	0.64

Note. mean score = 2.78; standard deviation scale = 0.65; Cronbach's Alpha α = 0.773; measurement was in German language.

Table 4 Interest in ecological sustainability topics (int.eco.sust.)

Item	Discrimination	M	SD	Difficulty
How much are you interested in...				
... climate change.	0.53	3.03	0.77	0.68
... natural disasters triggered by human action.	0.53	3.04	0.81	0.68

Note. mean score = 3.03; standard deviation scale = 0.69; Cronbach's Alpha α = 0.688; measurement was in German language.

Table 5 Motivation to act sustainably (mot.sust.)

Item	Discrimination	M	SD	Difficulty
For example, for topics like environmental protection, climate change or fair working conditions in developing countries, ...				
... I feel like dealing with them.	0.65	2.68	0.87	0.56
... I find that so important that I am very committed to it.	0.64	2.40	0.85	0.47
... I spend a lot of time to inform myself about this.	0.59	2.24	0.85	0.41
It is important for me to buy sustainably produced products.	0.49	2.80	0.83	0.60

Note. mean score = 2.53; standard deviation scale = 0.66; Cronbach's Alpha α = 0.787; CFI = 1; TLI = 1; RMSEA < .001; measurement was in German language.

3.3 Data Preparation and Analysis

The research question was answered by using structuring equation models (SEM) with both the outcome and the predictors as latent variables with multiple indicators. The performance in domain-specific requirement situations are assumed to be based primarily on underlying cognitive as well as affective-motivational dispositions (Blömeke et al. 2015). To test this assumption, the declarative knowledge scales as well as the questionnaire scales were used as predictors for SSKSM.

For the declarative knowledge tests (KBA, KSD, KSM) and the SSKSM, item parcels constructed used as indicators applying the same method as in Seeber et al. (2019). The parcel scores were treated as continuous variables. Response data from the questionnaire scales were analysed on the item level, treating the items as ordinal variables. All models were estimated using the lavaan package (Rosseel 2012) version 0.6–4 in R (R Development Core Team 2019). For all analyses including ordinal variables, the diagonally weighted least squares estimator was used. Missing responses were treated using the two-step procedure implemented in lavaan (Rosseel et al. 2019), allowing to use data from cases with incomplete data in the analyses. As measures for model fit, χ^2 values with corresponding degrees of freedom, the root mean squared error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis-index (TLI) are reported. Values of RMSEA $\geq .06$, CFI $\geq .95$, and TLI $\geq .95$ are considered as indicators for good model fit (Hu and Bentler 1999).

The analyses comprise three models: In model 1, a latent regression is analysed using the declarative knowledge dimensions as predictors for the SSKSM. Given the correlations reported by Seeber et al. (2019), we expect KSD to be the only significant predictor when all dimensions are used simultaneously. In a second step (model 2), the SSKSM, the significant predictors from the first model and the questionnaire scales are combined in one confirmatory factor analysis (CFA). The CFA provides (1) goodness of fit information for the model with all dimensions of interest combined and (2) latent correlations between all dimensions. In model 3, the research question is addressed by predicting competence in sustainability management by sustainability-related attitudes, motivation, and interests.

4 Results

In *model 1*, item parcels for the declarative knowledge tests were treated as parallel indicators (equal loadings, equal residual variances), the parcels for the SSKSM were treated as tau-equivalent (equal loadings). The model showed a good global

fit with $\chi^2 = 83.4$, $df = 62$, RMSEA = .036, CFA = .965, and TLI = .963. The good fit confirms the results of Seeber et al. (2019), who also confirmed the assumed structure of the four knowledge tests. Overall, 55.8% of the variance in competence in sustainability management was explained by the model, but only the regression weight for declarative knowledge about sustainable development from a societal perspective was significant (*standardized b* = .58, *SE* = .24, *p* = .013). Dropping the non-significant predictors from the model only marginally decreased the explained variance to 54.4%, the standardized regression weight of KSD alone is *b* = .74 (*SE* = .08, *p* < .001).

In *Model 2*, data from the SSKSM and KSD were combined with the questionnaire scales. The joint CFA model comprises seven dimensions: 1. SSKSM, 2. KSD, 3. ia.comp.res.sust., 4. aversion, 5. mot.sust., 6. int.chal.dev.count., and 7. int.eco.sust. Parcels for the tests were treated as in model 1. For all questionnaire items, loadings and residual variances were unrestricted. The model showed a good global fit with $\chi^2 = 521.70$, $df = 215$, RMSEA = .041, CFA = .981, and TLI = .978. Latent correlations are shown in Table 4. Both attitude dimensions (ia.comp.res.sust. & aversion) and mot.sust. are significantly correlated with SSKSM, while the two interest dimensions (int.chal.dev.count. and int.eco.sust.) are not. KSD has the highest correlation with SSKSM, and aversion has the highest correlation among the questionnaire dimensions.

Table 6 Latent correlations between all dimensions in the joint CFA

Dimension	(1)	(2)	(3)	(4)	(5)	(6)
(1) SSKSM	1.00					
(2) KSD	0.73	1.00				
(3) ia.comp.res.sust.	0.36	0.22	1.00			
(4) aversion	-0.47	-0.31	-0.65	1.00		
(5) mot.sust.	0.22	0.22	0.68	-0.42	1.00	
(6) int.chal.dev.count.	0.07	0.15	0.42	-0.22	0.71	1.00
(7) int.eco.sust.	0.15	0.31	0.55	-0.44	0.64	0.54

Note. All latent correlations except for the two greyed-out coefficients are significantly different from zero (*p* < .05).

SSKSM = performance of competence in sustainability management, KSD = declarative knowledge about sustainable development from a societal perspective, ia.comp.res.sust = individual attitude that companies have to take responsibility for sustainability, aversion = aversion to sustainability, mot.sust. = motivation to act sustainably, int.chal.dev.count. = interest in topics about challenges in developing countries, int.eco.sust = interest in ecological sustainability topics.

Model 2 is identical to model 1 regarding the measurement models. In the structural part, SSKSM is treated as dependent variable and all other dimensions as predictors. The model fit of model 2 is identical to the fit of model 1. Overall, 66.6 % of the variance in competence in sustainability management was explained by the model. Standardized regression weights for all predictors are presented in Table 4.

Table 7 Standardized regression coefficients (b) from the latent regression (Model 3) for SSKSM as outcome variable

Predictor	b	SE	P
KSD	0.701	0.099	0.000
ia.comp.res.sust.	0.194	0.176	0.135
Aversion	-0.253	0.133	0.029
mot.sust.	0.042	0.189	0.412
int.chal.dev.count.	-0.039	0.139	0.610
int.eco.sust.	-0.295	0.149	0.976

Note. *p*-values are based on one-sided tests for positive effects for all dimensions except for aversion to sustainability; SSKSM = performance of competence in sustainability management, KSD = declarative knowledge about sustainable development from a societal perspective, ia.comp.res.sust = individual attitude that companies have to take responsibility for sustainability, aversion = aversion to sustainability, mot.sust. = motivation to act sustainably, int.chal.dev.count. = interest in topics about challenges in developing countries, int.eco.sust = interest in ecological sustainability topics.

Apart from the significant effect of KSD, the only expected negative effect of aversion is significant, partially supporting H1. Effects of motivation and interests are non-significant, thus H2 and H3 have to be rejected. If the non-significant predictors are dropped from the model, only keeping KSD and aversion to sustainability, the explained variance decreases to 60.6%. Aversion to sustainability explains an increment of 6.3% variance above KSD alone. The incremental explained variance corresponds to an effect size of $f^2 = 0.16$, which can be classified as a medium effect size (Cohen 1992).

5 Discussion and Conclusion

Competences represent complex constructs that can be explained by a variety of cognitive and affective-motivational factors (Blömeke et al. 2015). Concerning the competence of sustainability management, only limited evidence toward the structural relationships between individual knowledge dimensions exists (Seeber et al. 2019). In this contribution, we follow conceptual and methodological assumptions by Seeber et al. (2019) for the analysis of competence in sustainability management. We reproduced findings from Seeber et al. (2019) regarding a 4-dimensional structure of the competence model (SSKSM, KBA, KSD & KSM). In addition, we expanded the research perspective from Seeber et al. (2019) by examining differentiated relationships and in particular the influence of affective-motivational dispositions on the performance of competence in sustainability management (SSKSM). Three hypotheses were developed, which are discussed below.

A first hypothesis assumes that sustainability-relevant attitudes have a positive influence on the performance of competence in sustainability. The results show a significant negative effect of the aversion on the SSKSM scale. Students with a higher aversion to sustainability develop weaker solutions in terms of sustainability management compared to students who reject the aversive items. The effect is substantial and in line with H1. Although a negative scale is used, the results are congruent with findings from studies in Vocational Education and Training (VET), which show a significant influence of sustainability-based attitudes on behavioural intentions to act sustainably at the workplace (Seeber and Michaelis 2014, Michaelis 2017).

The second attitude scale *ia.comp.res.sust* has no significant effect on the SSKSM scale. The average of this scale shows that high approval ratings are achieved on this scale; the scale has a slightly lower standard deviation than the scale aversion. Thus, the non-significant effect of the scale *ia.comp.res.sust* may possibly be attributed to its more general conception. We assumed that the approval toward the scale's items was easy for the students, while the importance of underlying values of these items would not consistently be considered in management situations of the SSKSM scale. The discrepancy between attitudes and their realisation in concrete situations is a well-known finding in sustainability research (Section 2).

The fact that H2 and H3 are to be rejected can also be related to the content of the scales. The motivation scale refers to a general perspective supporting a sustainable behaviour, which is a more general level of abstraction than to support sustainability management. We chose this level of abstraction as the majority of students have only weak comprehensive practical experience in sustainabil-

ity management. We expected that the motivation to act sustainably is also an indicator for the motivation to decide in management situations in the sense of sustainability management. However, no evidence exists that a motivation to act sustainably would be transferrable to professional domains such as sustainability management. Therefore, it is questionable whether the motivation to act sustainably is domain-specific. For example, due to some parallels to sustainability research, explanations could be found in moral research in VET. Contrary to theoretical assumptions, empirical findings show that morality is less a generic than a situation-specific construct (e.g., Beck et al. 2000, Minnameier 2011). In addition, the scale *mot.sust.* was deliberately phrased in the sense of an intrinsic motivation. Differentiated dimensions of motivation are not considered (like the assumptions by Ryan and Deci 2004). To sum up, the relationship between motivational dispositions and competence in sustainability management needs further research.

With regard to the scales of interest (*int.chal.dev.count.* & *int.eco.sust.*), the degree of abstraction should also be considered. These scales refer more to global challenges of a sustainable development (climate change, working conditions in developing countries). Here, the interest in sustainability management is not measured directly. Interest scales, which are more closely aligned with sustainability management, could produce a more pronounced effect.

In addition, the results indicate new insights into the structural relationship of the competence model for sustainability management. The KSD scale is the most important predictor of the SSKSM scale: The higher the declarative knowledge about sustainability from a societal perspective, the higher the coping with the requirements of sustainability management. Due to the significant effect of the scale aversion on SSKSM, however, an affective-motivational disposition is also relevant. These findings are in line with empirical analyses in VET (Seeber and Michaelis 2014, Michaelis 2017). However, the effect of the aversion scale is much less pronounced than the KSD scale could attribute to the theoretical background. Sustainability-related attitudes are subject to a large number of potential values (Shepherd et al. 2009), which were being considered in our analyses only exemplarily by two different attitude scales. We assume that even stronger effects can be measured taking into account a broader range of sustainability-related values.

In summary, sustainability-related attitudes can influence the performance of competence in sustainability management. The influence of the motivation to act sustainably as well as interest constructs on the performance of competence in sustainability management could not be confirmed. However, additional research considering more specific scales will be necessary. The analyses also demonstrate that declarative knowledge about sustainability from a societal perspective is the most important predictor for the performance of competence in sustainability

management. Reasons for the smaller effects of attitudes may also lie in a limited consideration of sustainability-relevant values in the attitude scales.

References

- Ajzen, I. (1989). Attitude structure and behavior. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (241–274). New York: Taylor & Francis.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2), 179–211.
- Bamberg, S., & Möser, G. (2007). Twenty years after Hines, Hungerford, and Tomera: A new meta-analysis of psycho-social determinants of pro-environmental behaviour. *Journal of environmental psychology*, 27(1), 14–25.
- Beck, K., Bienengraber, T., & Parche-Kawik, K. (2000). Entwicklungsbedingungen kaufmännischer Berufsmoral – Befunde zur beruflichen Primärsozialisation und Implikationen für die Weiterbildung. In C. Harteis, H. Heid, & S. Kraft (Eds.), *Kompodium Weiterbildung. Aspekte und Perspektiven betrieblicher Personal- und Organisationsentwicklung* (pp. 191–208). Opladen: Leske + Budrich.
- Biewen, M., Happ, R., Schmidt, S., & Zlatkin-Troitschanskaia, O. (2018). *Knowledge growth, academic beliefs and motivation of students in business and economics--A longitudinal German case study*. Higher Education Studies, 8(2), 9–28.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Brückner, S., & Pellegrino, J. W. (2017). Contributions of response processes analysis to the validation of an assessment of higher education students' competence in business and economics. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 31–52). Cham: Springer.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
- Gräsel, C., Bormann, I., Schütte, K., Trempler, K., Fischbach, R., & Asseburg, R. (2012). Perspektiven der Forschung im Bereich Bildung für nachhaltige Entwicklung. In BMBF (Eds.), *39 Bildung für nachhaltige Entwicklung—Beiträge der Bildungsforschung* (pp. 7–24). Berlin: BMBF.
- Happ, R., Förster, M., Zlatkin-Troitschanskaia, O., & Carstensen, V. (2016a). Assessing the previous economic knowledge of beginning students in Germany – implications for teaching economics in basic courses. *Citizenship, Social and Economics Education*, 15(1), 45–57.
- Happ, R., Zlatkin-Troitschanskaia, O., Beck, K., & Förster, M. (2016b). Increasing heterogeneity in students' prior economic content knowledge – Impact on and implications for teaching in higher education. In E. Wuttke, J. Seifried, & S. Schumann (Eds.), *Economic Competence and Financial Literacy of Young Adults* (pp. 193–210). Opladen: Barbara Budrich Publishers.
- Hartig, J., Klieme, E., & Leutner, D. (Eds.). (2008). *Assessment of competencies in educational contexts*. Göttingen: Hogrefe.

- Heckhausen, J., & Heckhausen, H. (2018). Motivation and action: Introduction and overview. In J. Heckhausen & H. Heckhausen (Eds.), *Motivation and action* (pp. 1–14). Cham: Springer.
- Hesselbarth, C., & Schaltegger, S. (2014). Educating change agents for sustainability—learnings from the first sustainability management master of business administration. *Journal of Cleaner Production*, 62, 24–36.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1), 1–55.
- Kaiser, F. G., Wölfling, S., & Fuhrer, U. (1999). Environmental attitude and ecological behaviour. *Journal of environmental psychology*, 19(1), 1–19.
- Leiserowitz, A. A., Kates, R. W., & Parris, T. M. (2006). Sustainability values, attitudes, and behaviors: A review of multinational and global trends. *Annual Review of Environment and Resources*, 31, 413–444.
- Lewalter, D., Wild, K. P., & Krapp, A. (2001). Interessenentwicklung in der beruflichen Ausbildung. In K. Beck & V. Krumm (Eds.), *Lehren und Lernen in der beruflichen Erstausbildung* (pp. 11–35). Opladen: Leske + Budrich.
- Lozano, R., Ceulemans, K., Alonso-Almeida, M., Huisingh, D., Lozano, F. J., Waas, T., ... & Hugé, J. (2015). A review of commitment and implementation of sustainable development in higher education: results from a worldwide survey. *Journal of Cleaner Production*, 108, 1–18.
- Lozano, R., Merrill, M., Sammalisto, K., Ceulemans, K., & Lozano, F. (2017). Connecting competences and pedagogical approaches for sustainable development in higher education. *A literature review and framework proposal. Sustainability*, 9(10), 1889.
- Michaelis, C. (2017). Kompetenzentwicklung zum nachhaltigen Wirtschaften: Eine Längsschnittstudie in der kaufmännischen Ausbildung. Frankfurt am Main: Peter Lang.
- R Development Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL <http://www.R-project.org>.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. URL <http://www.jstatsoft.org/v48/i02/>. [retrieved 12/07/2019]
- Rosseel, Y., Jorgensen, T. D., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, B., Scharf, F. (2019). *Package 'lavaan', July 3, 2019, version 0.6–4*. URL: <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf> [retrieved 12/07/2019]
- Ryan, R. M., & Deci, E. L. (2004). An overview of self-determination theory: An organismic-dialectical perspective. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3–33). Rochester: University Rochester Press.
- Schaltegger, S., & Synnestvedt, T. (2002). The link between ‘green’ and economic success: environmental management as the crucial trigger between environmental and economic performance. *Journal of environmental management*, 65(4), 339–346.
- Seeber, S., & Michaelis, C. (2014). *Development of a model of competencies required for sustainable economic performance among apprentices in business education*. Sig Workplace Learning, Paper Session, April 4, 2014, AERA Annual Meeting, Philadelphia/Pennsylvania, from April 3–7, 2014, downloadable in the AERA repository: <http://www.aera.net/repository>.

- Seeber, S., Michaelis, C., Repp, A., Hartig, J., Aichele, C., Schumann, S., Siepelmeyer, D., (2019). Assessment of competences in sustainability management: Analyses to the construct dimensionality. *Zeitschrift für Pädagogische Psychologie*, 33(2), 148–158.
- Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2005). Windows into the mind. *Higher education*, 49(4), 413–430.
- Shepherd, D. A., Kuskova, V., & Patzelt, H. (2009). Measuring the values that underlie sustainable development: The development of a valid scale. *Journal of Economic Psychology*, 30(2), 246–256.
- United Nations (n.d.). *Transforming our world: The 2030 Agenda for sustainable development*. URL: [https://sustainabledevelopment.un.org/content/documents/21252030 %20Agenda%20for%20Sustainable%20Development%20web.pdf](https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf)
- Wiek, A., Withycombe, L., & Redman, C. L. (2011). Key competencies in sustainability: a reference framework for academic program development. *Sustainability Science*, 6(2), 203–218.

Appendix

Table A.1 Matrix for the 2-dimensional scale of interest

Item	Factor 1	Factor 2	Communality
How much are you interested in...			
... economic problems in developing countries.	.93	-.10	.80
... ecological problems in developing countries (e.g. effects of monocultures).	.64	.23	.59
... social problems in developing countries (e.g. working conditions).	.66	.13	.53
... climate change.	-.04	.81	.63
... natural disasters triggered by human action.	.07	.73	.58



4.3

The Impact of Entry Preconditions on Student Dropout and Subject Change in Business and Economics

Kühling-Thees, C., Happ, R., Zlatkin-Troitschanskaia, O., and Pant, H. A.

Abstract

The proportion of students dropping out of higher education economics is remarkably high at approximately 25 % and has been constant for years. The results on the entry preconditions of dropping out and changing subjects in Bachelor's degree programs in business and economics presented in this paper are based on a representative longitudinal study in Germany. In two survey waves, the cognitive entry preconditions and characteristics of business and economics students related to their study and learning processes were tested. *Dropouts* have lower levels of previous knowledge than students *changing subjects* and students who *continue to study business and economics*. The students' characteristics related to study and learning processes also show significant differences between these three groups. Of particular relevance are the credit points achieved within the first academic year. The identified characteristics of dropouts and subject changers offer valuable implications for further higher education research and practice.

Keywords

Student dropout, subject change, entry preconditions, higher education economics, longitudinal study

Acknowledgements and Funding Information

The study was funded by the German Federal Ministry of Education and Research with the funding number 01PK15013. We would like to thank the anonymous reviewers for their constructive and valuable feedback.

1 Background and Research Questions

Over the past 15 years since the Bologna Reform (e.g., Hericks 2018; Albuquerque et al. 2019), there has been a substantial increase in the number of freshman students enrolling in business and economic (B&E) degree programs in Germany. While the number of freshman students in B&E has almost doubled within 10 years¹, the dropout rate in this domain is also high at 25 % (Heublein and Schmelzer 2018). There are several national and international studies on dropout of B&E degree programs and its influencing factors (e.g., Bosshardt 2004; Arnold 2013; Kercher 2018). Research on study success indicates the crucial importance of students' individual characteristics at the beginning of studies as dropout influencing factors (Arnold and Straten 2012; Belloc et al. 2011; Kuh et al. 2006; Schneider and Preckel 2017). The first academic year is considered to be particularly critical, as the majority of students² drop out of the course during this period (Barefoot 2004). Current research identifies cognitive abilities, such as previous school performance in the form of poor high school grades in specific school subjects or students' low general intellectual abilities, as particular risk factors for dropping out of university. Learning process characteristics such as academic achievements in the first academic year (e.g., number of successfully completed examinations, achieved credit

- 1 See the freshman student statistics for winter semester 2006/2007 (285, 851) and winter semester 2016/2017 (439, 369) (Federal Statistical Office 2007, 2017).
- 2 In Germany, the higher education system consists of different types of higher education institutions. The most common institutions are universities and universities of applied sciences (Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany 2017).

points etc.) are also regarded as factors conducive to dropout (Montmarquette et al. 2001; Bennet 2009; Belloc et al. 2011).

Despite this intensive research there is no consensus on the definition of dropout. Depending on how dropout is defined and operationalized, the underlying sample changes and the results of the analyses may differ (Grau-Valldosera and Minguillón 2014). Current research shows that the differentiation between various groups of dropout students provides added value in the analysis of dropouts. The different (sub)groups of dropouts can be compared with each other and specific reasons for dropping out can be identified (Trautwein and Bosse 2017; Rump et al. 2017). Therefore, for the present study a theoretical differentiation of dropout was established (Section 2.1).

Based on the current state of research, we examine the influencing factors for student dropouts in the domain of B&E using data from a current nationwide representative study (Zlatkin-Troitschanskaia et al. 2019). The longitudinal study allows for analysis of the students' individual characteristics at the beginning of B&E studies ("entry characteristics")³. In addition to previous education (such as a completed major B&E course in high school or vocational education), the freshman students' study-relevant previous knowledge as well as general intellectual ability (GIA) were measured using validated test instruments⁴. Based on this database, this paper presents findings on the following research questions (RQ):

1. How do the groups of dropouts, subject changers and continuing students differ from each other in terms of cognitive preconditions (prior knowledge, GIA)?
2. How do the groups of dropouts, subject changers and continuing students differ in terms of study- and learning process-related characteristics (number of exams taken, achieved credit points)?

In Section 2, the theoretical background and the state of research on dropout and changing subjects in the B&E domain are described in depth. Section 3 presents the study design and sample. Section 4 describes the empirical findings on the

3 The study also includes students of social science. The Social Science and B&E courses are based on different curricular principles and research traditions. The respective freshman students differ accordingly. This article focuses on the B&E course. For analyses on the dropout of students in social science, see Kühling-Thees, Happ, Zlatkin-Troitschanskaia, Schmidt, Jitomirski & Schlax (in review).

4 As the state of research in Section 2.2 shows, no other study for students in B&E could be found that includes these two entry preconditions of the test takers in the longitudinal analysis using validated tests.

two *RQs*. The paper concludes with limitations of the study and implications for further research (Section 5).

2 Theoretical Background

2.1 Conceptualization of Dropout and Subject Change

The literature on dropouts contains various definitions of dropout, which focus on different specific aspects (Grau-Valldosera and Minguillón 2014). The concept of dropout is often based on the non-achievement of a degree (European Commission 2015). According to this definition, students who have completed a degree in another subject after changing subject are not considered to be dropouts (Kauffmann 2015). In the research on study success, the focus lies on other indicators such as the completion of studies within the regular period of study (Svanum and Bigatti 2009), grades achieved in the final exams or success rates in exams (York and Gibson 2015; van der Zanden et al. 2018). In most studies, the dropout rate is the key factor in operationalizing study failure (Grau-Valldosera and Minguillón 2014; Ortiz and Dehon 2013). Here, four groups can be distinguished, which are often summarized under the term dropouts:

- a) *Dropouts*: According to Heublein (2014), these include students who started their first degree course but left the university system without a degree. When identifying this group, it should be noted that re-entry could take place at a later point in time and that individuals from this group could therefore return to higher education (Blüthmann 2012).
- b) *Subject changers*: In contrast to the dropouts, subject changers remain in the higher education sector (BMBF 2008). The decision for a change of subject can have many reasons. For example, the chosen subject was not the desired study course or was not perceived as suitable during the course of study (Vogel et al. 2018). A lack of study success or positive motivational experiences can also cause a change of subject.
- c) *University changers*: Several students in Germany retain their subject of study but change university over the course of their studies (Heublein 2014). This category refers to students who, for example, transferred to another university before completing their bachelor's degree and, accordingly, not those who transitioned from their bachelor's degree to their master's degree (Seemann 2015). Reifenberg et al. (2015) describe this group as "pure university changers".

- d) *University and subject changers*: Some students change both their study subject and university. According to the study by Reifenberg et al. (2015), this group accounts for the majority of changers (72%).

Previous studies point out the importance of considering these groups separately from each other (Blüthmann 2012). The reasons for deciding either to drop out or to change the subject differ. In the case of a change of higher education institution as well as a change of subject, there is a change within tertiary education, but in the case of only the second, there is a change in the subject domain (Blüthmann 2012). Dropping out of university is related to leaving tertiary education and is often caused by a number of different factors (Heublein 2014).

The present study is based on a dataset in which freshman students were surveyed at the beginning of their studies and after one academic year (end of the 2nd semester). The study thus focuses on dropout or change of subject in the particularly critical first year. Since the present study aims to identify the factors that contribute to the explanation of the reasons *for dropping out or changing the course of B&E study domain*, the underlying sample is differentiated into three groups. The first group includes the dropouts, the second the subject changers as well as subject-and-university changers and the third group the continuing students. In this study, a link to study success is established by regarding continuing students as the group of “successful” students. This group includes all students who continue to study B&E after the first year of study. The group of pure university changers is added to the group of continuing students, since there was no dropout or change from the B&E study domain.

2.2 State of Research on Dropout from B&E Studies

There are some studies on dropout from the B&E study domain and its major influencing factors, which provide indications of the process character of a study dropout (Heublein 2014), so that a longitudinal design is required for its explanation. In addition to the entry preconditions, the first year of study is particularly critical for the decision to drop out or change subject (Wendt et al. 2016). A dropout from B&E cannot be attributed to a few reasons; rather, a number of factors can be identified that influence a dropout.

Looking at cognitive determinants, studies indicate positive correlations between GIA and performance (Busato et al. 2000; Frey and Detterman 2004; Rohde and Thompson 2007). A lower intellectual ability increases the risk of dropping out of the study program (Heilbrun 1965). General cognitive skills are mentioned

among those skills that students are expected to possess for B&E studies. These abilities are usually operationalized by the grade of the higher education entrance qualification (German GPA equivalent) (Blüthmann 2012). GPA is considered the best individual factor for predicting academic success (Busato et al. 2000; Rohde and Thompson 2007; Yousef 2011). Beginners with a worse German GPA equivalent are more likely to terminate their studies early without a degree or to change subjects (Larsen et al. 2013; Smith and Naylor 2001).

Moreover, mathematical skills are of particular importance for B&E studies (Ballard and Johnson 2004; Markle 2017), as this study track places high mathematical demands (e.g., Arnold and Rowaan 2014). The risk of dropout due to deficits in a student's previous mathematical education increases (Arnold and Rowaan 2014; Arnold and Straten 2012; Ballard and Johnson 2004). A high level of mathematical skills has a positive effect on the performance in B&E studies; poorer mathematics skills have a negative effect on B&E studies and are associated with a higher risk of dropout (Anderson et al. 1994; Arnold and Rowaan 2014; Arnold and Straten 2012; Ballard and Johnson 2004).

Subject-specific previous knowledge is another important factor influencing student dropout (e.g., Arnold and Rowaan 2014). Previous knowledge in the subject represents the knowledge that is connectable to integrate new knowledge into the knowledge network (e.g., Sternberg and Ruzgis 1994). Every second dropout reports content-related overload during his/her studies if there is a lack of prior knowledge with less ability to connect. In the B&E domain, previous knowledge has a positive influence on study success (Happ et al. 2016; Brand and Xie 2010; Rienties et al. 2012).

Despite this research, it should be critically emphasized that not even a single study could be viewed which assessed student characteristics such as (B&E) previous knowledge or GIA using a validated test at the start of the course study, and thus before the decision to drop out or change the subject could have been made. Furthermore, no study could be found that had a longitudinal design with two measurement points. In contrast, in the majority of the studies, only a retrospective assessment of the reasons for the dropout or change of subject was made.

Based on the state of research, our study assesses both the cognitive entry preconditions of students using validated and reliable tests as well as the study- and learning process-related characteristics within a longitudinal design. O'Connor and Paunonen (2007) recommend to consider the different study performance-related criteria separately from each other. The first year's academic performance was identified as a relevant factor in the performance criteria. Grades (academic performance) have an influence on the decision to drop out; the worse the percep-

tion of one's academic achievement, the greater the tendency to drop out (Montmarquette et al. 2001; Bennet 2009; Stinebrickner and Stinebrickner 2014). Further results on study process characteristics indicate that study progress is to be considered as an indicator for dropout and thus further study- and learning process-related characteristics are also considered in more depth in this study (Babcock and Marks 2011; Scott-Clayton 2012).

3 Study Design and Sample

At the beginning of winter semester 2016/17, 7,679 B&E freshman students at 46 universities and universities of applied science were surveyed in a nationwide representative field study in paper-pencil design on their entry study preconditions (Zlatkin-Troitschanskaia et al. 2019). The study used a quasi-experimental design with representative random sampling of higher education institutions.⁵ The surveys took place at the participating universities as part of introductory lectures planned for freshman students. Trained test leaders carried out the survey; participation in the survey was voluntary and was rewarded with an incentive of €5.

The survey included the test-based assessment of study-relevant previous knowledge using items from the German adaptation of the fourth version of the Test of Economic Literacy⁶ (TEL4, for the US-American original, see Walstad et al. 2013; for the German adaption, see Happ et al. 2018) and the German adaptation of the fourth version of the Test of Understanding College Economics⁷ (TUCE4; for the US-American original, see Walstad et al. 2007; for the German adaptation, see Zlatkin-Troitschanskaia et al. 2014). On the basis of the validation studies (Happ et al. 2018), a total of 25 items in multiple-choice format were selected from the two tests to measure economic knowledge and understanding: 15 items from the TEL4 and 10 items from the TUCE4. One point was assigned for each correctly answered item, so that the total score can range from 0 to 25 points. The test shows a reliability of *Cronbach's* $\alpha = 0.72$.

To measure GIA, a short form of the Berlin Test for Fluid and Crystalline Intelligence (BEFKI) was used. This test captures figural reasoning (Schipolowski

5 Due to the challenging field access at the universities, it is associated with a high panel mortality rate (Heublein et al. 2014).

6 The TEL4 consists of two versions A and B with 45 items each. Both versions are connected with 10 anchor items, which are identical (Walstad et al. 2013).

7 The TUCE4 consists of two questionnaire parts A and B with 30 items each, which are not connected via anchor items, so that 60 items are available (Walstad et al. 2007).

et al. 2017). The short form of the BEFKI consists of 16 items in multiple-choice format with a maximum of 16 points to be achieved. For each item there are two subitems and the item is considered to be solved correctly if both subitems are correct. The reliability analysis showed a *Cronbach's α* of 0.67. In addition to the scale from the BEFKI, the German GPA-equivalent can also be used as an indicator of GIA (Veenman et al. 2005). Since B&E studies also require particular skills in mathematics (Ballard and Johnson 2004; Laging and Voßkamp 2017; Markle 2017), the mathematics school grade was also included. In addition, socio-demographic characteristics such as the parents' educational background, age, gender and family language were also assessed.

Since studies have shown the first year of study to be particularly critical for dropping out of studies or changing subjects (Section 2.2), the freshman students from the winter semester 2016/17 (t_1) were surveyed again at the end of summer semester 2017 (t_2). For this purpose, an address panel was set up in the first survey round in winter semester 2016/17 and the participants in the study were surveyed online at the end of summer semester 2017 with a further comprehensive questionnaire on study- and learning process-related characteristics and conditions at the university. The second survey included 1,236 participants. On the basis of this survey, three groups could be classified: dropouts, subjectchangers, and students who continue studying in B&E.

In the second survey study and learning process-related characteristics relevant to the course of study, such as time spent studying and academic performance (such as examinations passed, grades, achieved credit points; Costa et al. 2018) were surveyed.

Due to the longitudinal design and the linking of the two samples from t_1 and t_2 , we have data for 790 students on the study entry preconditions, which were already collected from the participants at the beginning of their studies, and not retrospectively, as is the case in many studies (Section 2.2).⁸ However, a notable number of students from t_1 could not be surveyed again at the second measurement, which indicates a bias in the sampling (for limitations, see Section 5).

In t_2 , 43 participants stated that they had left the university system (*group of dropouts*). 1,131 subjects are students at the same or a different university, who continue studying in the B&E subject selected at t_1 (*group of continuing students*). 121 participants have changed B&E for a different subject (*group of subject*

8 At the first measurement, approx. 35% provided an email address for the address panel. There are marginal differences between the respondents with and without the given email address (e.g. average age of the total sample of t_1 20.52 years; average age of the respondents without an email address 20.53 years).

changers) and are studying at a different or at the same university. The descriptive characteristics of the three groups are shown in Table 1.

Table 1 Sociodemographic characteristics

	Continuing students ($n^9 = 684$)	Subject changers ($n = 81$)	Dropouts ($n = 24$)	Total sample t_1 ($n = 7,679$)
Age	20.16 (2.41)	20.23 (2.45)	21.42 (3.45)	20.52 (2.75)
Gender				
Female	358 (52.26 %)	41 (50.62 %)	8 (33.33 %)	3,503 (45.67 %)
Male	327 (47.74 %)	40 (49.38 %)	16 (66.67 %)	4,168 (54.33 %)
Desired studies ¹	592 (86.93 %)	65 (80.25 %)	18 (75.00 %)	6,572 (86.58 %)
Migration background ¹	169 (24.67 %)	25 (30.86 %)	6 (25.00 %)	2,287 (29.84 %)
Family language other than German ¹	167 (24.38 %)	25 (30.86 %)	5 (20.83 %)	259 (3.4 %)
Educational background of parents				
Low ²	48 (7.14 %)	5 (6.25 %)	4 (16.66 %)	595 (7.85 %)
Medium ³	292 (43.45 %)	30 (37.50 %)	12 (50.00 %)	3,417 (45.11 %)
High ⁴	332 (49.41 %)	45 (56.25 %)	8 (33.34 %)	3,563 (47.03 %)

Note. SD in brackets. ¹ = yes. ² = no graduation or lower secondary school. ³ = secondary school or high school. ⁴ = study or doctorate.

In the overall sample of B&E freshman students ($n = 7,679$) surveyed in winter semester 2016/17 at t_1 , the average age was 20.50 years ($SD = 2.80$)¹⁰. The later dropouts were on average one year older (21.42 years, $SD = 3.45$) than the continuing students (20.16 years, $SD = 2.41$). In the total sample of 7,679 students, 45.22 % were female. The group of dropouts comprised more male students than female. In the other two groups, the gender ratio is almost equal.

With regard to the assessment of t_1 as to whether the B&E degree program is the desired course of study, as expected, fewer students from the groups of both sub-

9 Due to faulty panel codes for several test takers, not all study participants could be matched across t_1 and t_2 .

10 The comparison of individual characteristics of the overall sample at t_1 with the matched sample at t_2 shows significant differences (for limitations, see Section 5).

ject changers and dropouts (the fewest in this case) reported that the B&E degree program was their desired course of study. Since these data were already collected before the beginning of studies, no experiences made during the course of the studies could influence these ratings.

4 Results

4.1 Study Entry Preconditions

Regarding RQ1, the cognitive study-related entry preconditions differ across the sample (Table 2). In the group of dropouts, the average German GPA equivalent¹¹ was 2.53, in the group of subject changers 2.26 and in the group of continuing students 2.15. The difference between the groups is significant ($F(2, 776) = 6.11, p < .01$). The Bonferroni test shows a significant difference ($p < .01$) between the group of dropouts and the continuing students with an intermediate effect strength (Cohen's $d = 0.658$) (Table 3). The average grade in mathematics was 2.61 in the group of dropouts, 2.34 in the group of subject changers and 2.24 in the group of continuing students. The continuing students thus achieved both a better mathematics grade at t_1 and a better German GPA equivalent in comparison with the subject changers and dropouts. However, the difference between the groups is not significant ($F(2, 713) = 1.60, p > .05$).

At the beginning of winter semester 2016/17, the overall score in the test on B&E knowledge and understanding was 14.04 score points for the group of dropouts, 14.42 points for subject changers and 14.76 points for continuing students (max. score: 25 points). The difference between the groups is not significant ($F(2, 787) = 0.52, p > .05$). The average overall score in the test for GIA (BEFKI) was 9.34 (max. score: 16 points) for continuing students. In comparison, the subject changers achieved 8.48 points and the dropouts 8.71. The difference between the groups is significant ($F(2, 787) = 4.02, p < 0.01$). Further analyses show that there is a significant difference ($p = .02$) between the group of subject changers and the continuing students (Cohen's $d = 0.312$) (Table 3).

11 In Germany, the grade at the end of high school is important for the transition into higher education. Most German universities select their students based on the German equivalent to the high school GPA, which ranges from 1 (best grade) to 5 (worst grade).

Table 2 Overview of the cognitive entry preconditions before studying

	Continuing students (<i>n</i> = 685)	Subject changers (<i>n</i> = 81)	Dropouts (<i>n</i> = 24)	ANOVA
German GPA equivalent	2.15 (0.58) ¹	2.26 (0.53)	2.53 (0.51)	F(2, 776) = 6.11, <i>p</i> < .01
Grade mathematics	2.24 (0.93) ²	2.34 (0.89) ³	2.61 (1.02) ⁴	F(2, 713) = 1.60, <i>p</i> > .05
Previous knowledge in economics (max. 25 points)	14.76 (4.36)	14.42 (3.85)	14.04 (2.76)	F(2, 787) = 0.52, <i>p</i> > .05
General intellectual ability (max. 16 points)	9.34 (2.74)	8.48 (2.89)	8.71 (2.49)	F(2, 787) = 4.02, <i>p</i> < .01

Note. SD in brackets. ¹*n* = 674; ²*n* = 627; ³*n* = 72; ⁴*n* = 17

Table 3 Post-Hoc Bonferroni: Comparison between the groups for cognitive entry preconditions before studying

Comparison	Difference	SE	95% Confidence interval	
			Lower limit	Upper limit
GPA				
Subject changers vs. Dropouts	-0.28	0.13	-0.60	0.04
Continuing students vs. changers	-0.11	0.07	-0.27	-0.10
Continuing students vs. dropouts	-0.38**	0.12	-0.28	-0.10
General intellectual ability				
Subject changers vs. Dropouts	-0.23	0.64	-1.76	1.31
Continuing students vs. dropouts	0.64	0.57	-0.74	2.01
Continuing students vs. changers	0.86*	0.32	0.09	1.64

Note. **p* = .05, ***p* = .01, ****p* = .001

In summary, less favorable cognitive entry preconditions can be observed in the group of dropouts with regard to almost all the characteristics examined. However, the differences to the group of subject changers or continuing students are, in part, only slight or not significant.

4.2 Study- and Learning Process-Related Characteristics

Regarding RQ2, the study- and learning process-related characteristics significant differ between the groups. As Table 4 shows, subject changers invest the most time in the preparation and follow-up for courses (9.61 h/week; dropouts with 7.96 h/week and continuing students with 8.32 h/week). The same distribution is found in examination preparation (subject changers with 12.34 h/week, continuing students with 11.86 h/week and dropouts with 9.44 h/week). Dropout students invest the least time in both the preparation and follow-up for courses (7.96 h/week) and examination preparation (9.44 h/week). The differences between the groups are not significant (preparation and follow-up: $F(2, 1,156) = 1.65, p > 0.05$; examination preparation: $F(2, 1,152) = 1.29, p > 0.05$).

Table 4 Study- and learning process-related characteristics

	Continuing students ($n = 1,044$)	Subject Changers ($n = 88$)	Dropouts ($n = 27$)	ANOVA
Time spent in hours per week				
Preparation and follow-up courses	8.32 (6.51)	9.61 (6.93)	7.96 (5.87)	$F(2, 1,156) = 1.65, p > .05$
Attendance of courses	17.25 (6.44) ¹	17.53 (5.68) ²	13.90 (7.70)	$F(2, 1,152) = 3.75, p < .05$
Exam preparation	11.86 (8.25) ³	12.34 (8.90)	9.44 (7.83)	$F(2, 1,152) = 1.29, p > .05$
Taken exams	4.56 (1.56) ⁴	3.93 (1.87) ⁵	3.32 (2.06) ⁶	$F(2, 620) = 9.86, p < .001$
Passed exams	4.27 (1.77) ⁷	3.45 (2.53) ⁸	2.28 (1.90) ⁹	$F(2, 701) = 18.76, p < .001$
Credit points achieved so far	24.84 (14.52) ¹⁰	18.43 (9.99) ¹¹	14.12 (9.97) ¹²	$F(2, 682) = 12.93, p < .01$

Note. SD in brackets. ¹ $n = 1,041$; ² $n = 87$; ³ $n = 1,040$; ⁴ $n = 540$; ⁵ $n = 61$; ⁶ $n = 22$; ⁷ $n = 608$; ⁸ $n = 71$; ⁹ $n = 25$; ¹⁰ $n = 590$; ¹¹ $n = 69$; ¹² $n = 26$

In all three groups, the largest time expenditure is attributed to the attendance of courses, and is on a comparatively high level for the groups of subject changers and continuing students. Significant differences between the groups can be found ($F(2, 1,152) = 1.75, p < 0.05$). The group of subject changers differs significantly from the group of dropouts ($p = .03$; *Cohen's d* = 0.586) and the group of dropouts from the group of continuing students ($p = .02$; *Cohen's d* = 0.519) (Table 5). Accordingly, the differences are mainly in the time spent on-site at the university (intra-university expenditure of time) and less in the preparation and follow-up of the lectures and exam preparation (non-university expenditure of time).

The group differences in non-university time expenditure are also reflected in the exams taken or passed (Table 4). In comparison to subject changers, dropouts took fewer exams and passed fewer. Continuing students took and passed more exams than subject changers. The difference between the three groups is significant (for taken: $F(2, 620) = 9.86, p < 0.001$; for passed: $F(2, 701) = 18.76, p < 0.001$). The group of continuing students differs significantly from the group of dropouts as well as from the group of subject changers with low to medium effect strengths (subject changers: $p = 0.01$, *Cohen's d* = 0.395; dropouts: $p = 0.001$, *Cohen's d* = 0.784) (Table 5). With regard to exams passed, a significant difference between all three groups can be seen (continuing students – subject changers: $p = 0.001$; *Cohen's d* = 0.44; continuing students – dropouts: $p = 0.000$, *Cohen's d* = 1.121; changer – dropouts: $p = 0.021$, *Cohen's d* = 0.491) (Table 5).

Similar findings were found between the groups with regard to credit points (CP) achieved after the first semester (Table 4). The dropouts reached on average 14.12, subject changers 18.43 and continuing students 24.84 CP. The difference between the groups is significant ($F(2, 682) = 12.93, p < 0.001$). Remarkably, on average all groups have not reached the recommended amount of 30 CP after the first semester (e.g., European Commission 2015). The group of continuing students differed significantly from the group of subject changers ($p = 0.001$, *Cohen's d* = 0.454) and from the group of dropouts ($p = 0.000$, *Cohen's d* = 0.746) (Table 5).

Table 5 Post-Hoc Bonferroni: Comparison between groups for study- and learning process-related characteristics

Comparison	Difference	SE	95 % Confidence interval	
			Lower limit	Upper limit
Time spent in hours per week				
Attendance of courses				
Subject changers vs. dropouts	3.64*	1.41	0.25	7.03
Continuing students vs. dropouts	3.36*	1.25	0.37	6.36
Continuing students vs. changers	-0.28	0.72	-1.99	1.44
Taken exams				
Subject changers vs. dropouts	0.62	0.40	-0.35	1.58
Continuing students vs. changers	0.63**	0.22	0.11	1.15
Continuing students vs. dropouts	1.24***	0.35	0.40	2.09
Passes exams				
Subject changers vs. dropouts	1.17*	0.43	0.13	2.21
Continuing students vs. changers	0.82***	0.23	1.08	2.91
Continuing students vs. dropouts	2.00***	0.38	0.26	1.38
Credit points achieved				
Subject changers vs. dropouts	4.32	3.22	-3.40	12.04
Continuing students vs. changers	6.40***	1.78	2.13	10.67
Continuing students vs. dropouts	10.72***	3.83	4.00	17.45

Note. * $p = .05$, ** $p = .01$, *** $p = .001$

5 Conclusion, Limitations and Implications for Further Research

The results of this study indicate that dropout and change of subject in higher education are multicausal and complex processes. Particularly in the group of subject changers, fewer performance-related factors are decisive for the change of subject. In comparison, the group of dropouts is characterized by rather unfavorable entry preconditions.

Overall, the findings are in line with the current state of research (Wendt et al. 2016) and underline the importance of cognitive entry preconditions for studying B&E. The group of continuing students is characterized by higher levels of all assessed characteristics. This group shows a better German GPA equivalent and better GIA than the other two groups. Research on academic success suggests that the entry preconditions in particular are of high importance for academic success (e.g., Kuh et al., 2006). The subject-related prior knowledge acquired, for example,

through learning opportunities at school (e.g., B&E as a major subject at school) or extracurricular learning processes (e.g., students learning about economic content via mass and social media, see also Jitomirski et al. in this volume), is an important factor for the acquisition of knowledge over the course of studies. This can be well justified with theoretical principles of knowledge acquisition (e.g., Alexander 1997). Previous knowledge in these theories is characterized as connectable knowledge, which is systematically enhanced in the learning process through the expansion of the knowledge network. A more highly developed intellectual ability, which is also expressed in our study, proves to be beneficial for this.

The study also found significant differences between the groups in terms of study- and learning process-related characteristics, for instance, in the time spent within the university. Here, the study underlines the high importance of teaching and learning processes in higher education institutions. The findings presented here indicate that the learning opportunities on site, interaction with teachers and with fellow students are process characteristics that could reduce the probability of dropping out of studies. Continuing students also take the most exams and pass them more often. Especially the passed exams show significant differences between all three groups. The same findings can also be observed in the credit points achieved. Performance-related deficits can be shown particularly with regard to dropouts. In the current drop-out research, the study and learning process characteristics are still poorly and only partially examined (e.g., Babcock and Marks 2011; Scott-Clayton 2012). The analyses presented in this article demonstrate that these characteristics should be taken into account more differentiatedly.

A closer view on national and international research confirms that the present study is the first study in which cognitive entry preconditions were assessed using a validated tests and study- and learning process-related characteristics were examined in a longitudinal design. Thus, this study makes a unique and significant contribution to the research on dropout in the B&E domain and the presented findings provide an important basis for further research.

The longitudinal study design contributes to the high internal and external validity of the findings. One year later, approximately 16% of the total sample from t_1 of the sample could be reached again in t_2 due to the online-based field access. Of the study participants who agreed in t_1 to participate in t_2 as well (35.97%), 44.75% of the study participants who agreed in t_1 were able to participate in t_2 again. Of these, just under 15% no longer study in the B&E domain. Though loss of participants between t_1 and t_2 due to panel mortality and limited access to the field reduces the representativeness of the findings on t_2 , this study provides first meaningful indications as to the reasons for dropping out or changing subjects in this domain. However, the findings of the study must also be viewed critically in light of this sample selection. The overall sample at t_1 differs significantly from the

matched sample at t_2 regarding sociobiographical characteristics. This indicates a selection effect, which may also be due to the specific field of research. Dropping out of university is a personal experience of life that students may not want to be reminded of (e.g. emotional reasons). Furthermore, the accessibility of dropouts and subject changers is often generally restricted, since, for example, students may have stated their university email in the address panel instead of their personal email; in this case, the test takers can no longer be reached after they have left university since the given email address is no longer valid (e.g. de Leeuw and Lugtig 2015). As is common in low-stakes assessments, the panel mortality is higher than in high-stakes assessments (Biasi et al. 2018).

The database does not allow for the analysis of possible cohort effects that could cause differences between students on the basis of other social and environmental factors (Baltes et al. 1978). Moreover, even theoretically expected compositional effects cannot be analyzed (Kunter and Trautwein 2003), although the learning group structure among students could have a meaningful influence (Rump et al. 2017). Additionally, the analyses of study- and learning process-related characteristics for t_2 are based on the self-reported data of students, which may be biased and should therefore be critically evaluated. For the validation of the findings, the information on the actually achieved academic performances such as passed exams and grades would be of interest. However, such analyses would hardly be possible under German data protection law.

Further research also requires in-depth analyses of the curricular and intra-structural factors in the B&E study domain. The high importance of the time students spend at universities emphasizes that factors and processes such as interaction with teacher and other students should be given stronger research focus in further studies. In this paper, a strong focus was put on the analysis of cognitive entry preconditions. In future analyses, non-cognitive factors such as interest in the subject or intrinsic and extrinsic motivation for the chosen subject should also be included in the analyses.

References

- Albuquerque, C. P., Seixas, A. M., Oliveira, A. L., Ferreira, A. G., Paixão, M. P., & Paixão, R. P. (2019). *Higher education after Bologna: challenges and perspectives*. Coimbra: Coimbra University Press.
- Alexander, P. A. (1997). Mapping the multidimensional nature of domain learning: The interplay of cognitive, motivational, and strategic forces. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 213–250). Greenwich, CT: JAI Press.

- Anderson, G., Dwayne, B. & Fuss, A.M. (1994). The Determinants of Success in University Introductory Economic Courses. *The Journal of Economic Education*, 25 (2), 99–119.
- Arnold, I. J., & Straten, J. T. (2012). Motivation and Math Skills as Determinants of First Year Performance in Economics. *The Journal of Economic Education*, 43(1), 33–47.
- Arnold, I. J. (2013). Ethnic minority dropout in economics. *Journal of Further and Higher Education*, 37(3), 297–320.
- Arnold, I. J., & Rowaan, W. (2014). First-year study success in economics and econometrics: The role of gender, motivation, and math skills. *The Journal of Economic Education*, 45(1), 25–35.
- Babcock, P., & Marks, M. (2011). The Falling Time Cost of College: Evidence from Half a Century of Time Use Data. *The Review of Economics and Statistics*, 93(2), 468–478.
- Ballard, C. L., & Johnson, M. F. (2004). Basic math skills and performance in an introductory economics class. *The Journal of Economic Education*, 35(1), 3–23.
- Baltes, P. B., Cornelius, S. W., & Nesselroade, J. R. (1978). Cohort Effects in Behavioral Development: Theoretical and Methodological Perspectives. In: W. Andrew Collins (Eds.), *Minnesota Symposia on Child Psychology* (pp. 1–63). Lawrence Erlbaum: Hillsdale.
- Barefoot, B. O. (2004). Higher education's revolving door: Confronting the problem of student drop out in US colleges and universities. *Open Learning. The Journal of Open, Distance and e-Learning*, 19(1), 9–18.
- Bello, F., Maruotti, A., & Petrella, L. (2011). How individual characteristics affect university students drop-out: a semiparametric mixed-effects model for an Italian case study. *Journal of Applied Statistics*, 38(10), 2225–2239.
- Bennet, R. (2009). Determinants of Undergraduate Student Drop Out Rates in a University Business Studies Department. *Journal of Further and Higher Education*, 27(2), 123–141.
- Biasi, V., De Vincenzo, C., & Patrizi, N. (2018). Cognitive Strategies, Motivation to Learning, Levels of Wellbeing and Risk of Drop-out: An Empirical Longitudinal Study for Qualifying Ongoing University Guidance Services. *Journal of Educational and Social Research*, 8(2), 79–91.
- Blüthmann, I. (2012). *Studierbarkeit, Studienzufriedenheit und Studienabbruch: Analysen von Bedingungsfaktoren in den Bachelorstudiengängen* [Studyability, study satisfaction and drop-out: analyses of condition factors in bachelor's degree programmes] (Unveröffentlichte Dissertation, Freie Universität Berlin).
- Bosshardt, W. (2004). Student drops and failure in principles courses. *The Journal of Economic Education*, 35(2), 111–128.
- Brand, J. E., & Xie, Y. (2010). Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education. *American Sociological Review*, 75(2), 273–302.
- Bundesministerium für Bildung und Forschung (BMBF). (2008). *Studiensituation und studentische Orientierungen. 10. Studierendensurvey an Universitäten und Fachhochschulen* [Study situation and student orientations. 10th Student Survey at Universities and Universities of Applied Sciences]. Bonn, Berlin: BMBF.
- Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences*, 29(6), 1057–1068.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum.

- Costa, F. J. José da, Bispo, M. de S., & Pereira, R. de C. d. F. (2018). Dropout and retention of undergraduate students in management: a study at a Brazilian Federal University. *RAUSP Management Journal*, 53, 74–85.
- de Leeuw, E. D., & Lugtig, P. (2015). Dropouts in Longitudinal Surveys. *Statistics Reference Online*, 1–6.
- European Commission (2015). *Dropout and Completion in Higher Education in Europe*. Luxembourg: Publications Office of the European Union.
- Federal Statistical Office (2007). *Studierende an Hochschulen. Wintersemester 2006/2007* (Fachserie 11 Reihe 4.1) [Students at universities. Winter semester 2006/2007]. Wiesbaden: Statistisches Bundesamt.
- Federal Statistical Office (2017). *Studierende an Hochschulen. Wintersemester 2016/2017* (Fachserie 11 Reihe 4.1) [Students at universities. Winter semester 2016/2017]. Wiesbaden: Statistisches Bundesamt.
- Frey, M. C., & Detterman, D. K. (2004). Scholastic Assessment or g? The Relationship Between the Scholastic Assessment Test and General Cognitive Ability. *Psychological Science*, 15(6), 373–378.
- Grau-Valldosera, J., & Minguillón J. (2014). Rethinking Dropout in Online Higher Education: The Case of the Universitat Oberta De Catalunya. *International Review of Research in Open and Distance Learning*, 15(1), 290–308.
- Happ, R., Förster, M., Zlatkin-Troitschanskaia, O., & Carstensen, V. (2016). Assessing the previous economic knowledge of beginning students in Germany – implications for teaching economics in basic courses. *Citizenship, Social and Economics Education*, 15(1), 45–57.
- Happ, R., Zlatkin-Troitschanskaia, O., & Förster, M. (2018). How Prior Economic Education Influences Beginning University Students' Knowledge of Economics. *Empirical Research in Vocational Education and Training*, 10(5), 1–20.
- Heilbrun, A. B., Jr. (1965). Personality factors in college dropout. *Journal of Applied Psychology*, 49(1), 1–7.
- Hericks, N. (Ed.). (2018). *Hochschulen im Spannungsfeld der Bologna-Reform: Erfolge und ungewollte Nebenfolgen aus interdisziplinärer Perspektive [Universities at the intersection of the Bologna reforms]*. Wiesbaden: Springer.
- Heublein, U. (2014). Student Drop-out from German Higher Education Institutions. *European Journal of Education*, 49(4), 497–513.
- Heublein, U., & Schmelzer, R. (2018). *Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen. Berechnungen auf Basis des Absolventenjahrgangs 2016* (DZHW-Projektbericht) [The development of dropout rates at German universities. Calculations based on the 2016 graduate year]. Hannover: DZHW.
- Kauffman, H. (2015). A review of predictive factors of student success in and satisfaction with online learning. *Research in Learning Technology*, 23.
- Kercher, J. (2018). *Academic success and dropout among international students in Germany and other major host countries*. German Academic Exchange Service.
- Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006). *What matters to student success: A review of the Literature. Commissioned report for the National Postsecondary Education Cooperative*. Washington, DC: NPEC.
- Kunter, M., & Trautwein, U. (2003). *Psychologie des Unterrichts [Psychology of teaching]*. Paderborn: Verlag F. Schöningh.

- Kühling-Thees, C., Happ, R., Zlatkin-Troitschanskaia, O., Schmidt, U., Jitomirski, J., & Schlux, J. (in review). *Eine repräsentative Längsschnittstudie zum Studienabbruch und Fachwechsel in den Sozialwissenschaften* [A representative longitudinal study on the dropout and change of subject in the social sciences].
- Laging, A. & Voßkamp, R. (2017). Determinants of Maths Performance of First-Year Business Administration and Economics Students. *International Journal of Research in Undergraduate Mathematics education*, 3(1), 108–142.
- Larsen, M. R., Sommersel, H. B., & Larsen, M. S. (2013). *Evidence on Dropout Phenomena at Universities*. Copenhagen: Danish Clearinghouse for Educational Research.
- Markle, G. (2017). Factors influencing achievement in undergraduate social science research methods courses: A mixed methods analysis. *Teaching Sociology*, 45(2), 105–115.
- Montmarquette, C., Mahseredjian, S., & Houle, R. (2001). The Determinants of University Dropouts: A Bivariate Probability Model with Sample Selection. *Economics of Education Review*, 20(5), 475–484.
- O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of postsecondary academic performance. *Personality and Individual Differences*, 43(5), 971–990.
- Ortiz, E. A., & Dehon, C. (2013). Roads to success in the Belgian French community's higher education system: predictors of dropout and degree completion at the Université Libre de Bruxelles. *Research in Higher Education*, 54(6), 693–723.
- Reifenberg, D., Jörissen, J., & Peters, D. (2015). Ausgewählte Ergebnisse einer kooperativen Studie zu Hochschulwechsel und Studienabbruch [Selected Results of a Cooperative Study on University Change and Dropout]. *Zeitschrift für Qualitätsentwicklung in Forschung, Studium und Administration*, 9(3+4), 99–105.
- Rienties, B., Beusaert, S., Grohnert, T., Niemantsverdriet, S., & Kommers, P. (2012). Understanding academic performance of international students: the role of ethnicity, academic and social integration. *Higher Education*, 63(6), 685–700.
- Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, 35(1), 83–92.
- Rump, M., Esdar, W., & Wild, E. (2017). Individual differences in the effects of academic motivation on higher education students' intention to drop out. *European Journal of Higher Education*, 7(4), 341–355.
- Schipolowski, S., Wilhelm, O., & Schroeders, U. (2017). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz ab der 11. Jahrgangsstufe (BEFKI 11+)* [Berlin test for the assessment of fluid and crystalline intelligence from the 11th grade onwards]. Göttingen: Hogrefe.
- Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, 143(6), 565–600.
- Scott-Clayton, J. (2012). What Explains Trends in Labour Supply Among U.S. Undergraduates?. *National Tax Journal*, 65(1), 181–210.
- Secretariat of the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany (2017). *Basic Structure of the Education System in the Federal Republic of Germany*. Bonn: KMK.
- Seemann, W. (2015). Studienabbruch und Studienfachwechsel: Eine Studie zu den mathematisch-naturwissenschaftlichen Bachelorstudiengängen der Humboldt-Universität zu Berlin [Dropping out and changing subjects: A study on the mathematical-scientific

- bachelor courses at Humboldt-Universität zu Berlin]. *Zeitschrift für Qualitätsentwicklung in Forschung, Studium und Administration*, 9(3+4), 99–105.
- Smith, J., & Naylor, R. (2001). Determinants of degree performance in UK universities: A statistical analysis of the 1993 student cohort. *Oxford Bulletin of Economics and Statistics*, 63(1), 29–60.
- Sternberg, R. J. & Ruzgis, P. (Eds.). (1994). *Personality and intelligence*. New York: Cambridge University Press.
- Stinebrickner, T., & Stinebrickner, R. (2014). Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model. *Journal of Labor Economics*, 32(3), 601–644.
- Svanum, S., & Bigatti, S. M. (2009). Academic course engagement during one semester forecasts college success: Engaged students are more likely to earn a degree, do it faster, and do it better. *Journal of College Student Development*, 50(1), 120–132.
- Trautwein, C., & Bosse, E. (2017). The first year in higher education – critical requirements from the student perspective. *Higher Education*, 73(3), 371–387.
- Van der Zanden, P. J., Denessen, E., Cillessen, A. H., & Meijer, P. C. (2018). Domains and predictors of first-year student success: A systematic review. *Educational Research Review*, 23, 57–77.
- Veenman, M.V.J., Kok, R. & Blöte, A.W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science*, 33(3), 193–211.
- Vogel, C., Hochberg, J., Hackstein, S., Bockschecker, A., Bastiaens, T. J., & Baumöl, U. (2018). Dropout in Distance Education and how to Prevent it. In EdMedia+ *Innovate Learning* (pp. 1788–1799). Association for the Advancement of Computing in Education (AACE).
- Walstad, W. B., Rebeck, K., & Butters, R. B. (2013). *Test of economic literacy: Examiner's manual* (4th ed.). New York: Council for Economic Education.
- Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of understanding in college economics: Examiner's manual* (4th ed.). New York: National Council on Economic Education.
- Wendt, C., Rathmann, A., & Pohlenz, P. (2016). Erwartungshaltungen Studierender im ersten Semester: Implikationen für die Studieneingangsphase [Students' expectations in the first semester: implications for the introductory phase of studies]. In T. Brahm, T. Jenert & D. Euler (Eds.), *Pädagogische Hochschulentwicklung* (pp. 221–237). Wiesbaden: Springer Fachmedien.
- York, T. & Gibson, C.E. (2015). *Defining and Measuring Academic Success. Practical Assessment, Research & Evaluation*, 20(5), 1–20.
- Yousef, D. A. (2011). Academic Performance of Business Students in Quantitative Courses: A Study in the Faculty of Business and Economics at the UAE University. *Journal of Innovative education*, 9(2), 255–267.
- Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., Förster, M., & Brückner, S. (2019). Validating a Test for Measuring Knowledge and Understanding of Economics Among University Students. *Zeitschrift für Pädagogische Psychologie*, 33(2), 119–133.
- Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from a German Assessment of Business and Economics Competence. In H. Coates (Eds.), *Higher Education Learning Outcomes Assessment: International Perspectives* (pp. 175–197). Frankfurt am Main: Lang.



4.4

Influences on the Development of Economic Knowledge over the First Academic Year

Results of a Germany-Wide Longitudinal Study

Schlax, J., Zlatkin-Troitschanskaia, O., Kühling-Thees, C., and Brückner, S.

Abstract

Despite significant research, it remains unclear whether the goal of developing domain-specific knowledge in higher education is actually being achieved. This is also true for the internationally most popular study domain of business and economics. In Germany, a test for measuring economic knowledge was validated, enabling the analysis of change in knowledge over the course of studies. Business and economics students from across Germany were surveyed over the course of one study year: 7,111 beginning students in the winter term of 2016/2017, and 1,705 third semester students in the winter term of 2017/2018. Investigating the longitudinal matched sample of 734 students who took part at both measurement points, we found that economic knowledge developed slightly positively in the first year of study in economics. Prior economic knowledge, general intellectual ability and the courses attended are among the most important influencing factors of the knowledge test performance and grades after one academic year.

Keywords

Higher education, knowledge development, prior knowledge, general intellectual ability, economic knowledge test, longitudinal study, large-scale assessment

Acknowledgements and Funding Information

The study was funded by the German Federal Ministry of Education and Research with the funding number 01PK15001A.

We would like to thank the anonymous reviewers for their constructive and valuable feedback.

1 Development of Professional Knowledge in Higher Education Economics

The development of domain-specific competences according to academic and professional standards, and in particular the acquisition of content knowledge, is a central goal in higher education (Europe 2020 Strategy, EC 2010). Business and Economics (B&E) represents one of the most popular study domains both in Germany and internationally (Federal Statistical Office 2017; OECD 2018), as major and minor subjects (Federal Statistical Office 2017). To develop appropriate and effective instruction and to promote domain-specific knowledge in higher education, the targeted student learning outcomes and corresponding competencies need to be validly assessed (e.g., NRC 2012). The investigation of knowledge development requires longitudinal analyses, which are rarely implemented in higher education, in particular, in B&E (Zlatkin-Troitschanskaia et al. 2019a).

A knowledge assessment must meet the quality criteria of validity, reliability and objectivity (AERA, APA and NCME, 2014). This task is increasingly being addressed by studies which aim to assess economic competences (e.g., Schumann et al. 2013; Zlatkin-Troitschanskaia et al. 2014; Aprea and Wuttke 2016; Eberle et al. 2016), but so far only few of them have been longitudinal (Zlatkin-Troitschanskaia et al. 2013; Walstad and Rebeck 2008; Schmidt et al. 2016). Longitudinal studies are necessary to validly assess the development of domain-specific knowledge over the course of studies and to identify the central factors influencing the acquisition of knowledge (e.g., Shavelson et al. 2018). Only by identifying central influencing factors of domain-specific knowledge can help derive implications for the design of teaching practice and the appropriate support of student groups.

Various research has already identified significant influencing factors at the beginning of studies (Section 3), indicating that the first academic year is critical for the development of domain-specific knowledge (Kim and Lalancette 2013; Zlatkin-Troitschanskaia et al. 2017). The present nationwide longitudinal study with bachelor students of B&E based on the current state of research focuses on the following research questions (*RQ*):

1. How does economic knowledge develop over the first academic year (*RQ1*)?
2. Which factors influence the level of domain-specific knowledge after one academic year (*RQ2*)?

In addition to the objective assessment of domain-specific economic knowledge via a knowledge test, we also took into consideration the students' grades in B&E courses after one academic year (for grades as academic success criteria, see, e.g., Trapmann et al. 2007a) as additional indicators of study success and knowledge acquisition. In university practice, exams are used to assess knowledge or to measure learning progress and should therefore provide further indications on students' level of domain-specific knowledge of students after one academic year. In our study, we aim to clarify to what extent an added value can be expected when using a domain-specific knowledge test as a diagnostics instrument to assess students' content knowledge:

3. To what extent does the level of domain-specific knowledge at T2 correlate with university grades in B&E courses (*RQ3*)?
4. To what extent do the influencing factors considered in this study differ between the knowledge test score and university grades in B&E courses as two different indicators of acquisition of domain-specific knowledge (*RQ4*)?

After the conceptualization of economic domain-specific knowledge (Section 2), possible influencing factors are derived from the current state of research (Section 3), and investigated by regression analysis on the basis of the longitudinal sample (Section 4). The paper concludes with the interpretation and critical discussion of the results, and their implications for further research (Section 5).

2 Conceptualization of Domain-Specific Knowledge in B&E

Although there is a consensus that the aim of higher education is to develop professional competences, there are several different understandings and definitions of this term (e.g., Blömeke et al. 2015; Blömeke et al. 2013). Professional competence is generally understood as domain-specific knowledge and expertise (Ericsson and Smith 1991; Gruber and Mandl, 1996; Bromme 2001; Gruber 2001), constituting a cognitive performance (e.g., Klieme and Leutner 2006). Other definitions focus on action-related competence. Boyatzis (1982), for example, describes competence as personal capabilities that result in effective job performance. Spencer and Spencer (1993) stated: “A *competency* is an underlying characteristic of an individual that is causally related to criterion-referenced effective and/or superior performance in a job or situation. Underlying characteristic means the competency is a fairly deep and enduring part of a person’s personality. [...] Causally related means that a competency causes or predicts behavior and performance. Criterion-referenced means that the competency actually predicts who does something well or poorly, as measured on a specific criterion or standard” (p. 9). Especially renowned internationally is the conceptualization by Weinert (2001), who understands competence as learnable cognitive abilities and skills required to solve problems: „... the concept [competence] refers to the necessary prerequisites available to an individual or a group of individuals for successfully meeting complex demands.” (p. 62) and „... learning processes are a necessary condition for the acquisition of prerequisites for successful mastery of complex demands” (p. 63). Domain-specific knowledge can therefore be considered as an important part of domain-specific competences.

In the economic context, Macha and Schuhen (2012; see also the ECOS study) describe economic competence as the ability to recognize and describe economic issues with verbal or mathematical content, to draw appropriate inferences, and to apply the knowledge in a reflected way when solving economic problems. To conceptualize B&E knowledge, Zlatkin-Troitschanskaia and colleagues (2014) developed a three-dimensional model which differentiates between a content dimension of Business and Economics (e.g., finance, microeconomics), a cognitive dimension according to the teaching-learning taxonomies (e.g., remembering content, applying, evaluating), and a dimension related to the structure of knowledge (propositional, case-related and strategic; for further information, see Zlatkin-Troitschanskaia et al. 2014). The study presented here focuses on economic knowledge as a subfield of B&E competence (for a definition of economic competence, see also Beck 1989).

3 State of Research on the Development of Domain-Specific Knowledge over the Course of B&E Studies and Its Influencing Factors

To evaluate how domain-specific knowledge develops over the course of studies and which factors might influence it, the change in knowledge needs to be measured in a valid way (Happ et al. 2016; Schmidt et al. 2016). Since there was no validated instrument for measuring domain-specific knowledge at higher education level in economics and its change in German-speaking countries (Zlatkin-Troitschanskaia et al. 2017), the WiWiKom I project (Zlatkin-Troitschanskaia et al. 2014) and its follow-up project WiWiKom II (Zlatkin-Troitschanskaia et al. 2019) focused on the validation of a test to assess domain-specific knowledge in higher education economics according to the standards for Educational and Psychological Testing (AERA et al. 2014). This enables a differentiated examination of domain-specific knowledge over the course of studies as well as its potential influencing factors.

Some influencing factors such as prior knowledge (e.g., Happ et al. 2018) can be identified from previously existing research. Learning opportunities are also considered a significant factor in the acquisition of knowledge, although the existence of learning opportunities does not necessarily go hand in hand with their use (for the conceptual background and for factors influencing the use of learning opportunities and their impact on knowledge development, see, e.g., the Learning Opportunity-Use Model by Helmke and Schrader 2013). The present study focuses on prior knowledge, general intellectual ability and other student characteristics such as interest in B&E topics, as previous studies have shown that these are strong influencing factors. Based on the current state of research (for more details, see Zlatkin-Troitschanskaia et al. 2019) the following hypotheses (H) can be formulated:

H1: There is a difference between the level of economic knowledge at the beginning of studies and after one academic year, with a higher level of knowledge after one year.

H2: The level of economic knowledge after one academic year is related to the domain-specific courses attended over the first two study semesters.¹

Several studies indicate a positive influence of general intellectual ability (measured by school grades or psychological tests) on the development of domain-specific knowledge (Prins et al. 2006; Anderson et al. 1994; Trapmann et al. 2007b; Giese et al. 2013). Accordingly, a higher level of general intellectual ability also indicates a higher acquisition of knowledge:

1 In Germany, the academic year is not divided into trimesters but two semesters of 6 months each.

H3: The level of economic knowledge after one academic year is related to the general intellectual ability (measured by the grade of university entrance qualification (UEQ) and an intelligence test).

Further studies show the influence of prior learning opportunities, which are often assessed using indicators such as domain-specific learning opportunities completed as advanced courses at school or the completion of vocational training, on knowledge acquisition (Happ et al. 2018; Wuttke and Beck 2002; Schumann et al. 2013; Erdel 2010).

H4: The level of economic knowledge after one academic year is related to the prior domain-specific learning opportunities attended before the beginning of studies.

Motivational aspects and students' interest in B&E topics are also considered factors that influence the acquisition of knowledge, whereby a connection is also assumed between the two aspects (Krapp 1999). According to previous research findings, interest in study-related content and (learning) motivation have a positive effect on knowledge acquisition (Biewen et al. 2018; Erdel 2010).

H5: The level of economic knowledge after one academic year is related to students' interest in B&E topics at the beginning of studies.

In addition, sociodemographic factors such as gender, age and migration background have also been deemed important for economic knowledge development (for more details on the advantage of male students, see, e.g., Hasler and Lusardi 2017; Bucher-Koenen et al. 2017; Brückner et al. 2015; Jirjahn 2007; for the effect of age, see, e.g., Erdel 2010; Lammers et al. 2001; for effects of migration background, see, e.g., Happ et al. 2019; 2018). These are therefore included as control variables in the present study.

4 Methods

4.1 Study Design

The present study is based on data from two representative surveys of the WiWiKom II project (Zlatkin-Troitschanskaia et al. 2019). In the winter semester of 2016/2017 (T1), 7,111 bachelor students of B&E at 48 German universities were surveyed using a paper-pencil-based questionnaire. Universities with B&E degree programs were contacted, and surveys were carried out at the universities willing to participate in this study. The surveys took place in introductory lectures for beginning students. The surveys were conducted by test leaders trained in the standardized procedure of test administration. Participation in the surveys was voluntary and

incentivized with 5 €. One year later, in the winter semester of 2017/2018 (T2), 1,705 students at 22 universities were surveyed again in courses for second-year students. Surveys in T2 were carried out at the universities in which a sufficiently high number of students were assessed in T1, so that a sufficiently high matching rate could be expected. Participation in T2 was also voluntary and was rewarded with 10 €. The re-participating students were identified through individual pseudonymized panel codes.

4.2 Instruments

The questionnaire for T1 consisted of sociodemographic questions, the WiWiKom knowledge test in economics (with items from the TEL IV, Walstad et al. 2013, and TUCE IV, Walstad et al. 2007) and an intelligence test (BEFKI, Schipolowski et al. 2017). The testing time was about 45 minutes. The questionnaire in T2 consisted of sociodemographic questions, the knowledge test as well as questions about study progress, for instance, the courses attended during the first year of study.

Domain-specific knowledge after one academic year was assessed using two indicators: the knowledge test score and the grades of attended courses. The *domain-specific knowledge test* consists of 25 items from the US American Test of Economic Literacy (TEL IV; Walstad et al. 2013) and the Test of Understanding in College Economics (TUCE IV; Walstad et al. 2007), adapted for the German context (Zlatkin-Troitschanskaia et al. 2014, 2019a). The test covers the areas of fundamental economic knowledge, macro- and microeconomics. The multiple-choice items go beyond declarative knowledge since a deeper understanding of the contents and their application is necessary for responding to the item correctly. One point was assigned for each correctly answered item so that the total score could range from 0 to 25. As is frequently the case in knowledge assessment (Glug 2009; Baker and Kim 2004), missing answers were assumed to be the result of a lack of knowledge and were therefore considered as incorrect responses (for discussion, see Zlatkin-Troitschanskaia et al. 2019b; Schlax et al. in review). For missing values in more than 50 % of the items, the total score was set to missing, since in such cases, factors such as insufficient test motivation are to be assumed. The test was given in two orders, whereby the mean sum value did not differ between the questionnaire orders (T1: $x(\text{order } 1, n = 3,498) = 13.25 \pm 4.343$, $x(\text{order } 2, n = 3,532) = 13.28 \pm 4.434$, $p = 0.778$; T2: $x(\text{order } 1, n = 856) = 13.67 \pm 5.047$, $x(\text{order } 2, n = 846) = 13.72 \pm 5.00$, $p = 0.821$). In T1, the knowledge test achieved a reliability rating of *Cronbach's* $\alpha = 0.858$.

The *attended courses* over the first study year were assessed in T2 and divided into several study areas. In terms of subjects, courses in business, economic fundamentals, macro- and microeconomics were differentiated. The number of courses attended could be indicated as “none”, “1” or “2 or more”. The students’ grades² in the attended courses were also assessed in T2. If students attended several courses in one subject, they were asked to give the best grade.

The students’ *general intellectual ability* was operationalized by using a matrices test to measure fluid intelligence and the *grade of UEQ*. The UEQ grade represents the overall average grade and should be indicated as a decimal number. In the study, the subtest on fluid intelligence of the *Berlin Test of Fluid and Crystalline Intelligence* (BEFKI; Schipolowski et al. 2017) was used in T1. The test consists of 16 items, each consisting of two subitems with matrices. There are three possible solutions per subitem, one of which is correct. Only if both subitems were solved correctly, one point was assigned, so that the sum score could range from 0 to 16. The reliability analyses show a *Cronbach’s α* of 0.64.

In our study, the focus was on the measurement of fluid intelligence instead of other facets of general intelligence such as verbal intelligence, as logical reasoning is particularly important to understand complex models and phenomena in economics and to solve economics problems. Furthermore, fluid intelligence tests are considered to be well suited to avoid language-based bias, which was particularly important in this project, as evidence of language effects had already been found in previous studies (Brückner et al. 2015; Zlatkin-Troitschanskaia et al. 2019b).

Since the use of single indicators of general intellectual ability is often criticized in literature on this topic (e.g., Spiel et al. 2006; Hell et al. 2008), we included both indicators in our study (UEQ grade and the intelligence test), which also allows us to examine to what extent these two indicators contribute to explaining the variance of the dependent variables.

Due to the low correlation between the UEQ and the intelligence test ($r = -0.277$; grades are inverted in Germany), it can be assumed that the grade does not measure only fluid intelligence but rather measures another facet (verbal, crystalline) of general intelligence. With regard to the assumptions for linear models, there is still no multicollinearity if both factors are included (Variance Inflation Factor VIF < 10). These analyses suggest that the different indicators of general intelligence used in this study measure different constructs and can contribute to the explanation of dependent variables.

2 In the German grading system, grades are expressed numerically, with lower grades corresponding to better performance.

Prior *knowledge* was assessed using the knowledge test described above, which was also used in T1. The same test was used at both measurement times, without controlling for the order version (due to one-year interval). In addition, the completion of a commercial vocational training, the achievement of the UEQ at a school with an economic focus and the attendance of an advanced economics course in school were assessed.

Subject-related interest was assessed by using one item (“Are you interested in economics- and business-related topics?”). The assessed *socio-demographic variables* such as gender, age and migration background were included as control variables. Migration background was assessed using the question “Was at least one parent not born in Germany?” as well as questions about language skills. The participants were asked in which language they could communicate best (in German, in another language, in German just as well as in another language).

4.3 Sample

The participants were students of B&E (284/734, 38.56 %), B&E education (3/734, 0.41 %), business (256/734, 34.88 %), economics (101/734, 13.76 %), business engineering (67/734, 9.13 %) and business informatics (11/734, 1.50 %; 12/734, 1.63 % missing). The sample considered here refers to the group of participants who were in their first semester in T1 and also took part again in T2, i.e. students who actually form the longitudinal sample ($N = 734$). On the basis of this subsample, statements can be made about the change in domain-specific knowledge from the beginning of studies to the end of the first academic year (for a description of the longitudinal sample, see Table 1).

4.4 Statistical Methods

RQ1 is examined using a dependent t-test. Due to a low to medium high *ICC* of 0.095 and as the B&E education study programs differ between the participating universities, RQ2 is examined using a multi-level modeling (MLM) of the T2 knowledge score using the universities ($n = 22$) as a grouping variable (variance-constant) = 1.061 ± 0.480 , variance(residual) = 9.854 ± 0.631 ; LR vs. linear modeling: $\chi^2 = 33.65$, $p < 0.0001$).

As the *ICC* score of 0.047 (variance(constant) = 0.023 ± 0.017 , variance(residual) = 0.529 ± 0.035 ; LR vs. linear modelling: $\chi^2 = 5.74$, $p = 0.008$) regarding the grade in business courses and the *ICC* score for the grade in the economics

modules ($ICC = 0.038$) is only slightly below the common five percent limit (variance(constant) = 0.036 ± 0.035 , variance(residual) = 0.832 ± 0.065 ; LR vs. linear modeling: $\chi^2 = 2.27$, $p < 0.066$), MLM were also conducted. With regard to the grade achieved in the microeconomics modules, the grouping variable again showed a low to medium high variance proportion ($ICC = 0.073$; variance(constant) = 0.067 ± 0.056 , variance(residual) = 1.122 ± 0.087 ; LR vs. linear modeling: $\chi^2 = 3.31$, $p = 0.034$).

A random intercept model with explanatory variables on level 1 was modelled for each dependent variable. Although it can be assumed that the attended classes and their influence vary between universities, a corresponding random slope model did not lead to a significantly better model fit for any of the dependent variables (for all $p > 0.375$ in LR χ^2 test). The formula is as follows:

$$\text{dependent variable}_{ij} = \beta_0 + \beta_1 \text{independent variable1}_{ij} + \beta_2 \text{independent variable2}_{ij} + \dots + u_i + \varepsilon_{ij}$$

Differences in sample sizes between the multilevel models are due to missing values in the included variables. The deviance, AIC and BIC were calculated to indicate the model fit. In the significance test using the LR test, the models to be compared had to be adapted to the same sample size. The models were always compared with the next larger model with regard to the model fit.

With regard to the comparability of the measurement, the analyses showed a configurable measurement invariance of T1 and T2, indicating that the same construct was measured at both measurement points (CFI = 0.93, RMSEA = 0.018; χ^2 difference test $p < 0.001$). The program R (R Core Team, 2018) was used to perform the longitudinal match. The analyses were conducted using Stata Version 15 (Stata Corp, 2017).

Table 1 Descriptive statistics for the samples from T1, T2 and the matched sample from T1-T2

Variable	T1 (1st term) N = 7,111	T2 (3rd term) N = 1,705	Match T1 & T2 N = 734
Gender, male	3,841 (54.01 %; mis: 0.08 %)	933 (54.72 %; mis: 0.00 %)	365 (49.73 %; mis: 0.14 %)
Preferred communication language, German	6,839 (96.17 %; mis: 0.65 %)	1,648 (96.66 %; mis: 0.00 %)	712 (97.00 %; mis: 0.54 %)
Migration background, none	5,025 (70.67 %; mis: 0.14 %)	1,194 (70.03 %; mis: 0.00 %)	529 (72.07 %; mis: 0.00 %)
Advanced course in B&E, no	4,721 (66.39 %; mis: 0.49 %)	1,139 (66.80 %; mis: 0.23 %)	475 (64.71 %; mis: 0.14 %)
UEQ at school with B&E focus, no	5,179 (72.83 %; mis: 0.15 %)	n/a	587 (79.97 %; mis: 0.14 %)
B&E related vocational training, no	5,934 (83.45 %; mis: 0.15 %)	1,429 (83.81 %; mis: 0.00 %)	618 (84.20 %; mis: 0.00 %)
Age	20.41 (SD = 2.692; mis: 0.24 %)	21.12 (SD = 2.247; mis: 1.06 %)	19.91 (SD = 2.239; mis: 0.00 %)
UEQ grade	2.37 (SD = 0.569; mis: 2.33 %)	2.28 (SD = 0.583; mis: 1.06 %)	2.247 (SD = 0.576; mis: 0.68 %)
Score knowledge T1	13.27 (SD = 4.389; mis: 1.14 %)	n/a	14.13 (SD = 4.163; mis: 0.82 %)
Score BEFKI	8.26 (SD = 2.725; mis: 0.00 %)	n/a	8.86 (SD = 2.671; mis: 0.00 %)
Attended courses in first study year in business, at least one	n/a	1,604 (94.07 %; mis: 3.28 %)	693 (94.41 %; mis: 2.45 %)
Attended courses in first study year in economics, at least one	n/a	1,075 (63.05 %; mis: 15.31 %)	476 (64.85 %; mis: 12.94 %)
Attended courses in first study year in microeconomics, at least one	n/a	1,130 (66.28 %; mis: 11.03 %)	526 (71.66 %; mis: 7.77 %)
Attended courses in first study year in macroeconomics, at least one	n/a	554 (32.49 %; mis: 17.01 %)	233 (31.74 %; mis: 12.94 %)
Grade business	n/a	2.39 (SD = 0.893; mis: 8.97 %)	2.34 (SD = 0.902; mis: 7.77 %)
Grade economics	n/a	2.70 (SD = 1.010; mis: 41.29 %)	2.72 (SD = 1.050; mis: 38.42 %)
Grade microeconomics	n/a	2.79 (SD = 1.163; mis: 38.83 %)	2.87 (SD = 1.210; mis: 32.56 %)

Note. B&E= Business and Economics, UEQ = university entrance qualification, BEFKI = fluid intelligence short scale; in Germany, lower grades indicate better performance

5 Results

5.1 Difference between Both Measurements

The overall score in the economic knowledge test of the complete sample at the first measurement point (first semester) was 13.27 (± 4.389). At the second measurement point (third semesters), it was 13.69 (± 5.023) for the complete sample. The longitudinal sample showed a total change between T1 and T2, from $x(T1) = 14.12$ (± 4.163) to $x(T2) = 14.57$ (± 4.919). For *RQI (H1)*, the difference in knowledge level is significant ($t(727) = -3.321, p = 0.0013$) and can be interpreted as a small effect (*Cohen's d* = 0.095) according to Cohen (1988).

5.2 Relationship between Indicators of Economics Knowledge at T2

The grades in the domain-specific courses are considered not as predictors but as dependent variables, as they also represent indicators of study success and domain-specific knowledge (e.g., Trapmann et al. 2007a). As shown in Table 2, all dependent variables considered here cover different aspects of domain-specific knowledge, as the correlations are medium high.

Table 2 Correlations between indicators of domain-specific knowledge

	Sum score test	grade in business	grade in economics	grade in micro-economics
Sum score test	1.000			
grade in business	-0.320	1.000		
grade in economics	-0.262	0.735	1.000	
grade in micro-economics	-0.318	0.575	0.766	1.000

Note. Grades in Germany, i.e., the lower the number, the better the grade.

5.3 Factors Influencing Domain-Specific Knowledge

5.3.1 Learning Opportunities over the Course of Studies

The full model show that the number of business courses attended is of predictive importance with regard to the knowledge score after one academic year (Table 3). The other variables are not able to contribute to the explanation of variance of the test score.

The number of attended courses in the respective subject (with the exception of the grades in economics), also contributes significantly to explaining the variance of the grades (for business grade, attended courses in business $b = -0.434$, $SE = 0.091$, $p < 0.001^*$; for microeconomics grade, attended courses in microeconomics $b = -0.375^3$, $SE = 0.164$, $p = 0.022^*$; see appendix 1–3).

3 The negative sign is due to the inverse coding in the German grading system (lower numbers indicate better grades).

Table 3 MLM with knowledge test score at T2 as dependent variable (group variable: 22 universities)

Model: Wald $\chi^2 = 550.34, p < 0.001$				
Variable	b	SE	z	p
Constant	3.237	1.137	1.05	0.195
Attended courses in business	0.865	0.288	3.01	0.003*
Attended courses in economics	0.016	0.235	0.07	0.946
Attended courses in microeconomics	-0.137	0.286	-0.48	0.632
Attended courses in macroeconomics	.0,164	0,325	-0.50	0.614
UEQ grade	-1.608	0.301	-5.34	< 0.001*
Score BEFKI	0.128	0.572	2.24	0.025*
Knowledge test score at T1	0.605	0.042	14.56	< 0.001*
UEQ at school with B&E focus, no	0.796	0.387	2.06	0.040*
B&E related vocational training, no	0.394	0.446	0.88	0.377
Advanced course in B&E, no	-0.574	0.314	-1.83	0.067
Interest in B&E related topics	-0.011	0.227	-0.05	0.961
Gender, male	1.039	0.305	3.41	0.001*
Age	0.074	0.071	1.04	0.300
Migration background, none	0.394	0.341	1.16	0.247
Preferred communication language, German	.1.189	1.137	1.05	0.296

Note. $N = 557$ (due to missing values), UEQ = university entrance qualification, BEFKI = fluid intelligence short scale; in Germany, lower grades indicate better performance, * indicates significance on a 5 %-level.

5.3.2 General Intellectual Ability

Both predictors of general intellectual ability, the UEQ grade and the fluid intelligence score, significantly predict the knowledge score at T2, with the UEQ grade having the higher predictive power (Table 3).

When explaining the variance of the grades in the B&E courses, the UEQ grade has a significant predictive value for all courses (grade in business: $b = 0.586$, $SE = 0.070$, $p < 0.001$; grade in economics: $b = -0.007$, $SE = 0.020$, $p < 0.001^*$; grade in microeconomics: $b = 0.592$, $SE = 0.120$, $p < 0.001^*$), whereas the test score does not contribute significantly to explaining the variance (see appendix 1–3).

5.3.3 Prior Domain-Specific Knowledge

If all prior-knowledge-related variables are considered, only the knowledge score at the beginning of the study and visiting a school with B&E focus are significant with regard to predicting the level of knowledge after one academic year (Table 3). While the knowledge test score has a positive influence, to have visited a B&E-focused school does not.

The economics test score at T1 can also contribute significantly to explaining the variance of the grades in business courses ($b = -0.036$, $SE = 0.010$, $p < 0.001^*$), economics courses ($b = -0.049$, $SE = 0.014$, $p = 0.001^*$) and microeconomics courses ($b = -0.041$, $SE = 0.016$, $p = 0.012^*$).

Other indicators of prior knowledge do not have an influence on grades, except for the factor of vocational training attended, which influences business grades at university (note the inverse grading system; $b = 0.291$, $SE = 0.105$, $p = 0.006^*$; see appendix 1–3).

5.3.4 Interest in Study-Related Topics

The interest in B&E topics shows no significant predictive power regarding the domain-specific knowledge score after one academic year (Table 3). In the explanation of the variance of the grades in B&E courses after one year, interest again does not have a significant influence (grade in business: $p = 0.361$; grade in economics: $p = 0.594$; grade in microeconomics: $p = 0.733$; see appendix 1–3).

Table 4 Fit indices in multi-level models

Fit Indices	M0 (N = 729)	P	M1 (N = 566)	P	M2 (N = 562)	P	M3 (N = 557)	P	M4 (full; N = 557)
Knowledge score at T2									
Deviance	4,164.37	**	3,214.86	**	3,114.32	**	2,876.07	0.961	2,876.06
AIC	4,178.37		3,236.86		3,140.32		2,910.07		2,912.07
BIC	4,210.52		3,284.59		3,196.63		2,983.55		2,989.87

Note. Deviance = (-2)*log-likelihood-value. * = sig. 5 % level, ** = sig. on 1 % level. M0 = control variables, M1 = attended courses added, M2 = intellectual ability added, M3 = learning opportunities added, M4 = interest added.

Table 4 shows that the model fit for predicting the knowledge test score improves significantly with every model expansion, except for interest in B&E topics, where no significant improvement in terms of model fit could be determined.

Similar results can be found for the model fits regarding the grades. However, for the grade in the microeconomics module, no significant added value can be achieved with regard to the model fit by adding the attended courses to the model (see appendix 4).

6 Discussion

6.1 Summary and Interpretation of the Findings

In the matched sample presented here, a significant positive but only slight change in domain-specific knowledge was identified between the two measurement points (*RQ 1*). At the second measurement after one year of studying B&E, domain-specific knowledge is influenced by various factors (*RQ 2*).

In the multi-level model analysis, attending domain-specific courses, the UEQ grade and the domain-specific knowledge score at T1 were the main factors influencing the level of knowledge at T2 (measured via a test and via grades in domain-specific courses at university after one academic year). In particular, courses in business and microeconomics attended in the first year of studying B&E show a significant relevant influence on domain-specific knowledge.

With regard to general intellectual ability, the UEQ grade as an indicator was the most significant influencing factor. Furthermore, the level of domain-specific knowledge can be explained significantly by the students' level of prior knowledge at the beginning of their course of study in higher education. The score on the domain-specific knowledge test at T1 as an indicator of prior knowledge is the strongest predictor for the score on the knowledge test at T2.

When considering learning opportunities attended prior to studying in higher education as indicators of prior knowledge and their influence on the knowledge test score at T2, the results show that the learning opportunities seem to have a negative influence, or none at all. Students' interest in B&E topics at the beginning of their course of study can only explain the level of economic knowledge at T2 to a negligible extent.

Overall, the results confirm the influence of the factors general intellectual ability, learning opportunities attended over the course of study and prior knowledge. Domain-specific knowledge already acquired prior to beginning a course of study can explain the level of knowledge after one year to a significant extent, both in

terms of the domain-specific knowledge test score and the grades in the corresponding courses, even if other important influencing factors such as measures of intellectual ability and attended courses are taken into consideration.

The medium correlations between the domain-specific knowledge test and the grades in B&E-related courses indicate that the knowledge test can measure additional facets of knowledge after one academic year (RQ3). However, the central influencing factors show their influence on all knowledge level indicators (RQ4). The knowledge score at T1 contributes in all cases significantly to the explanation of variance. These findings support the incremental validity (e.g., Hunsley and Meyer 2003) of the economics knowledge test, and in accordance with further test validation results (Zlatkin-Troitschanskaia et al. 2019a, b), also supports the suitability of the test as an entry test. Administering this knowledge test at the beginning of B&E degree courses could make it possible to uncover students' potential need for support and to offer them effective measures through preliminary courses and tutorials before the beginning of their studies. Since courses in "mass subjects" such as B&E are hardly "adaptable" to individual needs, effective support for students at an early stage is of particular importance to achieve the teaching and learning goals for as many students as possible, even with a low degree of individualization (Zlatkin-Troitschanskaia et al. 2017, 2019a).

6.2 Limitations

One limitation of the present study and of low-stakes assessments (LSAs) in general is the proportion of study drop-outs, so that in this study only about one ninth of the initial sample has been reassessed in T2. Analyzing the descriptive values, the participants in the match sample are comparable to those in the original sample in T1, albeit slightly better-performing in terms of prior knowledge and measures of general intellectual ability. This indicates a slightly positive self-selection of the sample in T2. Due to voluntary participation despite the full survey design, sample distortions cannot be ruled out for T1 (e.g., non-participation and absence in surveyed lectures). Using computer-based handwriting comparison, we are currently trying to identify further matches between the participants in T1 and T2 that could not be identified by using the panel code.

A further disadvantage of LSA is the lower motivation of the participants to show maximum performance, as there are no significant consequences for the students (see also Hartig 2007, on test motivation; Biasi et al. 2018; Ramm et al. 2006). In this study, however, to motivate the participants, individual and group-comparative feedback on the test results was offered along with monetary incentives, which

the students could check after each measurement (Zlatkin-Troitschanskaia et al. 2019a). As the incentives were the same at every measurement point, significantly differing motivation effects are less likely.

Due to the multiple-choice format, guessing can also play an important role in the knowledge test (for adjusting approaches, see Smith and Wagner 2018). Although Walstad et al. (2018) could identify only a minimal significant effect using the same knowledge test in one US pre-post study, the possible differential guessing effects need to be analyzed in following studies.

As the same knowledge test was presented in both measurements, retesting effects at a one-year interval are unlikely, however they cannot be ruled out entirely (e.g., Happ et al. 2016). Further limitations of the measuring instruments used are the non-optimal internal consistency of the fluid intelligence test (for more details, see Jitomirski et al., in this volume) the assessing of socio-demographic and study-related variables including grades via self-reporting. In the courses, only the attendance of the courses was reported, but not the engagement or learned content, which may, however, have a higher significance for the level of knowledge than the attendance. The latter also applies to the assessment of attended learning opportunities prior to study.

Furthermore, when operationalizing economic knowledge, we are only capturing one part of the domain-specific knowledge taught in B&E (Zlatkin-Troitschanskaia et al. 2014). This is also indicated by the low correlations between the test scores and the grades after one academic year. As shown in previous projects, including tasks on additional content (sub)domains such as finance and accounting would increase the curricular and instructional validity of the knowledge test (e.g., Förster et al. 2015), but also reduce the usability and efficiency of the test in practical implementation.

Additionally, due to the limited test time, students' domain-specific interest was assessed with one question only, and only before they began their course of study. Future research should consider both a more differentiated operationalization and assessment of interest as well as the fact that subject-related interest may vary over the course of studies and that reciprocal relationships between the acquisition of domain-specific knowledge and grades can also be assumed.

Moreover, again due to the limited test time, only fluid intelligence could be assessed using an intelligence test but none of the other facets of general intelligence such as verbal intelligence, which might influence a student's performance in a text-based knowledge test.

Methodologically, multi-level modeling was conducted despite relatively small groups with only two-digit group sizes (Snijders and Bosker 2011) and without weighting, although the sub-samples of the universities were not the same size.

Despite the reported limitations, the longitudinal dataset is nationally and internationally unique, which is a significant strength of the study, and the already high informational value of the test results for the first academic year allow for important implications for further research in higher education and for higher education practice to be drawn.

6.3 Implications for Further Research

Although a significant part of the knowledge level after one academic year can already be explained in this study, there is still a high proportion of unexplained variance, especially with regard to grades, so that future research on further possible influencing factors is required. The included factors should be augmented by further indicators and tested using structural equation models to also analyze possible mediator effects, as independent effects of the individual factors on the dependent variable are theoretically and empirically questionable. Indeed, the paradoxical negative effect of pre-study learning opportunities can also be caused by the fact that their positive effect is already reflected in the variable of the test score at T1. A mediation effect can be assumed here as well. Additionally, specific institutional aspects should be addressed as well in further research.

The low knowledge growth in the first academic year requires more in-depth research. Preliminary work that takes a closer look at the development paths of knowledge (Happ et al. 2016; Schlax et al. 2019) suggests that, in addition to an increase in knowledge and the maintenance of prior knowledge, there is also a high degree of no knowledge change and loss of knowledge. Further research, including mixed method approaches, is intended to get to the root of the causes of knowledge development, such as the acquisition of misconceptions (Schmidt et al. 2020). The results of Schlax et al. (2019) also indicate that a high degree of repetition of already existing knowledge becomes evident in the first academic year. In-depth curricular analyses and cognitive interviews with the students could provide an explanation of the findings as shown in previous studies (e.g., Brückner and Pellegrino 2016).

Based on this study, analyzing knowledge development in the second and third years of bachelor programs is also required. The WiWiKom II project provides unique longitudinal data for this purpose, the analysis of which promises further insights to explain and promote the development of domain-specific knowledge over a course of B&E studies. The results to date already provide promising indications of the possible use of the test instrument in university practice (Zlatkin-Troitschanskaia et al. 2019a).

References

- American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME). (2014). *The standards for educational and psychological testing*. Washington DC: American Psychological Association.
- Anderson, G., Benjamin, D., & Fuss, M. A. (1994). The determinants of success in university introductory economics courses. *Journal of Economic Education*, 25(2), 99–119.
- Apra, C., & Wuttke, E. (2016). Financial Literacy of Adolescents and Young Adults: Setting the Course for a Competence-Oriented Assessment Instrument. In C. Apra, E. Wuttke, K. Breuer, N. Keng Koh, P. Davies, B. Greimel-Fuhrmann & J. Lopus (Eds.), *International Handbook of Financial Literacy* (pp. 397–414). Singapore: Springer.
- Baker, F.B., & Kim, S.H. (Eds.). (2004). *Item Response Theory: Parameter Estimation Techniques*. New York: Dekker.
- Beck, K. (1989). “Ökonomische Bildung”-Zur Anatomie eines wirtschaftspädagogischen Begriffs [“Economic Education”-On the Anatomy of an Economic Education Concept]. *Zeitschrift für Berufs- und Wirtschaftspädagogik [Journal for Business and Vocational Education]*, 85, 579–596.
- Biasi, V., De Vincenzo, C., & Patrizi, N. (2018). Cognitive Strategies, Motivation to Learning, Levels of Wellbeing and Risk of Drop-out: An Empirical Longitudinal Study for Qualifying Ongoing University Guidance Services. *Journal of Educational and Social Research*, 8(2), 79–91.
- Biewen, M., Happ, R., Schmidt, S., & Zlatkin-Troitschanskaia, O. (2018). Knowledge Growth, Academic Beliefs and Motivation of Students in Business and Economics – A longitudinal German Case Study. *Higher Education Studies* 8, 9–28.
- Blömeke, S., Gustafsson, J., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie [Journal for psychology]*, 223, 3–13.
- Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., & Fege, J. (2013). Modeling and Measuring Competencies in Higher Education. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and Measuring Competencies in Higher Education: Tasks and Challenges* (pp. 1–12). Rotterdam: Sense.
- Boyatzis, R. E. (1982). *The Competent Manager. A Model For Effective Performance*. New York: John Wiley & Sons, Inc.
- Bromme, R. (2001). Teacher Expertise. In N. J. Smelser, P. B. Baltes & F. E. Weinert (Eds.), *International Encyclopedia of the Behavioral Sciences: Education* (pp. 15459–15465). London: Pergamon.
- Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., & Walstad, W. B. (2015). Effects of prior economic education, native language, and gender on economic knowledge of first-year students in higher education. A comparative study between Germany and the USA. *Studies in Higher Education*, 40(3), (437–453).
- Brückner, S., & Pellegrino, J. W. (2016). Integrating the Analysis of Mental Operations into Multilevel Models to Validate an Assessment of Higher Education Students’ Competency in Business and Economics. *Journal of Educational Measurement*, 53(3), 293–312.
- Bucher-Koenen, T., Lusardi, A., Alessie, R., & Van Rooij, M. (2017). How financially literate are women? An overview and new insights. *Journal of Consumer Affairs*, 51(2), 255–283.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Routledge.
- Eberle, F., Schumann, S., Ackermann, N., Jüttler, A., & Kaufmann, E. (2016). *Modellierung und Messung wirtschaftsbürgerlicher Kompetenz (WBK), Valorisierungsbericht z.H. des SBFi* [Modelling and measurement of economic civic competence (WBK), valorisation report for the attention of the SBFi]. Accessed from <https://www.wiwi.uni-konstanz.de/schumann/forschung/forschungsprojekte/abgeschlossene-projekte/modellierung-und-messung-wirtschaftsbuergerlicher-kompetenz/> (17th June 2019).
- Erdel, B. (2010). *Welche Determinanten beeinflussen den Studienerfolg? Eine empirische Analyse zum Studienerfolg der ersten Kohorte der Bachelorstudenten in der Assessmentphase am Fachbereich Wirtschaftswissenschaften der Friedrich-Alexander-Universität Erlangen-Nürnberg* [Which determinants influence study success? An empirical analysis of the academic success of the first cohort of bachelor students in the assessment phase at the Department of Economics of the Friedrich-Alexander-University Erlangen-Nuremberg]. Accessed: <https://www.ssoar.info/ssoar/bitstream/handle/document/22022/ssoar-2010-erdel-welche-determinanten-beeinflussen-den-studienerfolg.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2010-erdel-welche-determinanten-beeinflussen-den-studienerfolg.pdf> (25th September 2015).
- Ericsson, K. A., & Smith, J. (1991). *Toward a general theory of expertise: Prospects and limits*. New York: Cambridge University Press.
- European Commission (EC) (2010). Europe 2020: A strategy for smart, sustainable and inclusive growth. Brussels: European Commission. Accessed: <http://ec.europa.eu/eu2020/pdf/COMPLETE%20EN%20BARROSO%20%20%20007%20-%20Europe%202020%20-%20EN%20version.pdf> (17th June 2019).
- Federal Statistical Office (Destatis) [Statistisches Bundesamt] (2017). Bildung und Kultur: Studierende an Hochschulen Fachserie 11 Reihe 4.1 [Education and Culture – Students at Universities – Preliminary Report Winter Term 2016/17 (subject series 11, series 4.1)]. Wiesbaden: Destatis Statistisches Bundesamt.
- Förster, M., Zlatkin-Troitschanskaia, O. & Happ, R. (2015). *Adapting and Validating the Test of Economic Literacy to Assess the Prior Economic Knowledge of First-Year Students in Business and Economic Studies in Germany* (Discussion Paper; Annual Meeting of the American Economic Association). Boston: AEA.
- Giese, S., Otte, F., Stoetzer, M. W. & Berger, C. (2013). Einflussfaktoren des Studienerfolges im betriebswirtschaftlichen Studium: Eine empirische Untersuchung [Influencing factors of study success in business studies: an empirical study]. *Jena Contributions to Economic Research*, 1.
- Glug, I. (2009). Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung [Development and validation of a multiple-choice test for the assessment of process-related scientific basic education]. Dissertation, Christian-Albrechts-Universität zu Kiel. Accessed: https://macau.uni-kiel.de/servlets/MCRFileNodeServlet/dissertation_derivate_00002755/dissertation_ingauglug.pdf (21st June 2019).
- Gruber H. & Mandl H. (1996). Expertise und Erfahrung [Expertise and experience]. In H. Gruber & A. Ziegler, (Eds.), *Expertiseforschung [Expertise research]* (pp. 18–34). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Gruber, H. (2001). Acquisition of expertise. In N. Smelser & P. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences*. Amsterdam: Elsevier.

- Happ, R., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2016). An Analysis of Economic Learning among Undergraduates in Introductory Economics Courses in Germany. *The Journal of Economic Education*, 47(4), 300–310.
- Happ, R., Zlatkin-Troitschanskaia, O., & Förster, M. (2018). How Prior Economic Education Influences Beginning University Students' Knowledge of Economics. *Empirical Research in Vocational Education and Training*, 10(5), 1–20.
- Happ, R., Nagel, M., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2019). How migration background affects master degree students' knowledge of business and economics. *Studies in Higher Education*, 1–16. doi: 10.1080/03075079.2019.1640670
- Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus [Scaling and definition of competence levels]. In: B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung [Language skills. Concepts and measurement]* (pp. 83–99). Weinheim: Beltz Verlag.
- Hasler, A., & Lusardi, A. (2017). The gender gap in financial literacy: A global perspective. Global Financial Literacy Excellence Center. Accessed: <https://gflec.org/wp-content/uploads/2017/07/The-Gender-Gap-in-Financial-Literacy-A-Global-Perspective-Report.pdf> (19th June 2019).
- Hell, B., Linsner, M. & Kurz, G. (2008). Prognose des Studienerfolgs [Prognosis of study success]. In M. Rentschler & H.-P. Voss (Eds.), *Studieneignung und Studierendenauswahl – Untersuchungen und Erfahrungsberichte [Aptitude for studies and student selection – Studies and experience reports]* (pp. 132–177). Aachen: Shaker.
- Helmke, A., & Schrader, F.-W. (2013). Angebots-Nutzungs-Modell [offer-utilization-model]. In M. A. Wirtz (Eds.), *Dorsch – Lexikon der Psychologie [Lexicon of Psychology]* (p. 147–148). Bern: Huber.
- Hunsley, J., & Meyer, G. J. (2003). The Incremental Validity of Psychological Testing and Assessment: Conceptual, Methodological, and Statistical Issues. *Psychological Assessment*, 15(4), 446–455.
- Jirjahn, U. (2007). Welche Faktoren beeinflussen den Erfolg im wirtschaftswissenschaftlichen Studium [Which factors influence the success of economics studies]? *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung [Schmalenbachs journal for economic research]*, 59(3), 286–313.
- Kim, H., & Lalancette, D. (2013). Literature Review on the Value-added Measurement in Higher Education. OECD. Accessed: <http://www.oecd.org/edu/skills-beyond-school/Literature%20Review%20VAM.pdf> (19th June 2019).
- Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG [Competence models for assessing individual learning outcomes and for balancing educational processes. Description of a newly established priority programme of the DFG]. *Zeitschrift für Pädagogik [Journal for Pedagogy]*, 52(6), 876–903.
- Krapp, A. (1999). Intrinsische Lernmotivation und Interesse. Forschungsansätze und konzeptuelle Überlegungen [Intrinsic learning motivation and interest. Research approaches and conceptual considerations]. *Zeitschrift für Pädagogik [Journal for Pedagogy]*, 45(3), 387–406.
- Lammers, W. J., Onweugbuzie, A. J., & Slate, J. R. (2001). Academic success as a function of gender, class, age, study habits, and employment of college students. *Research in the Schools*, 8(2), 71–81.

- Macha, K., & Schuhen, M. (2011). Modellierung ökonomischer Kompetenz in der Pilotstudie zu ECOS [Modelling economic competence in the ECOS pilot study]. In H. J. Schlösser & M. Schuhen, (Eds.), *Siegener Beiträge zur ökonomischen Bildung [Siegens Contributions to Economic Education]*. Siegen: ZöBiS.
- National Research Council: Committee on Defining Deeper Learning and 21st Century Skills, Pellegrino, J. W. & Hilton, M. L. (Eds.). Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National Academies Press.
- OECD (2018). *Education at a Glance 2017: OECD Indicators*. OECD Publishing. Accessed: <http://dx.doi.org/10.1787/eag-2017-en> (19th June 2019).
- Prins, F. J., Veenman, M. V. J., & Elshout, J. J. (2006). The impact of intellectual ability and metacognition on learning: New support for the thresholds of problematicity theory. *Learning and Instruction*, 16(4), 374–387.
- Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., & Schiefele, U. (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente [PISA 2003: Documentation of the survey instruments]*. Münster: Waxmann.
- R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Schipolowski, S., Wilhelm, O., & Schroeders, U. (2017). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz ab der 11. Jahrgangsstufe (BEFKI 11+) [Berlin test of fluid and crystallized intelligence for grades 11 and above]*. Göttingen: Hogrefe.
- Schlax, J., Zlatkin-Troitschanskaia, O., Schmidt, S., Kühling-Thees, C., Jitomirski, J. & Happ, R. (2019, April). *Analyzing Learning Processes and Distinct Learning Patterns in Higher Education Economics*. Paper presented at Annual Meeting of the American Educational Research Association, Toronto, Canada.
- Schlax, J., Zlatkin-Troitschanskaia, O., Happ, R., Pant, H. A., Jitomirski, J., Kühling-Thees, C., Förster, M., & Brückner, S. (in review). *Validity and Fairness of a New Entry Diagnostics Test in Higher Education Economics*. Manuscript submitted for publication.
- Schmidt, S., Brückner, S., Zlatkin-Troitschanskaia, O., & Förster, M. (2015). Das wirtschaftswissenschaftliche Wissen in der Hochschulbildung – eine Analyse der messinvarianten Erfassung finanzwirtschaftlichen Fachwissens bei Studierenden [The economic knowledge in higher education – an analysis of the measurement invariant assessment of financial economic expertise among students]. *Empirische Pädagogik [Empirical Pedagogy]*, 29(1), 106–124.
- Schmidt, S., Zlatkin-Troitschanskaia, O., & Fox, J.-P. (2016). Pretest-Posttest-Posttest Multilevel IRT Modeling of Competence Growth of Students in Higher Education in Germany. *Journal of Educational Measurement*, 53(3), 332–351.
- Schmidt, S., Zlatkin-Troitschanskaia, O. & Walstad, W. W. (2020). IRT Modeling of Decomposed Student Learning Patterns as Positive and Negative Learning in Higher Education Economics. In *Frontiers and Advances in Positive Learning in the Age of InformaTiOn (PLATO)*.
- Schumann, S., Eberle, F., & Oepke, M. (2013). Ökonomisches Wissen und Können am Ende der Sekundarstufe II: Effekte der Bildungsgang-, Klassen- und Geschlechtszugehörigkeit [Economic knowledge and skills at the end of upper secondary level II: effects of educational pathways, class and gender affiliation]. In U. Faßhauer, B. Fürstenau, & E. Wuttke

- (Eds.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung 2013* [Yearbook of Vocational and Business Education Research 2013] (pp. 35–46). Leverkusen: Budrich.
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Marino, J. (2018). Performance Indicators of Learning in Higher-Education Institutions: Overview of the Field. In E. Hazelkorn, H. Coates & A. Cormick (Eds.), *Research Handbook on Quality, Performance and Accountability in Higher Education* (pp. 249–263). Cheltenham: Edward Elgar.
- Smith, B. O., & Wagner, J. (2018). Adjusting for guessing and applying a statistical test to the disaggregation of value-added learning scores. *The Journal of Economic Education*, 49(4), 307–323.
- Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London: SAGE Publications.
- Spencer, L. M. & Spencer, S. M. (1993). *Competence at Work: Models for Superior Performance*. New York: Wiley.
- Spiel, C., Litztenberger, M., & Haiden, D. (2006). *Bildungswissenschaftliche und psychologische Aspekte von Auswahlverfahren* [Educational and psychological aspects of selection procedures]. University Wien. Retrieved from http://www.univie.ac.at/Psychologie/bildungspsychologie/download/auswahlverfahren_endbericht.pdf
- StataCorp (2017). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.
- Trapmann, S., Hell, B., Hirn, J.-O. W. & Schuler, H. (2007a). Meta-Analysis of the Relationship Between the Big Five and Academic Success at University. *Zeitschrift für Psychologie*, 215(2), 132–151.
- Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007b). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse [The validity of school grades for predicting study success – a meta-analysis]. *Zeitschrift für pädagogische Psychologie* [Journal for Educational Psychology], 21(1), 11–27.
- Walstad, W. B., Watts, M. & Rebeck, K. (2007). *Test of Understanding in College Economics: Examiner's manual* (4th ed.). New York: National Council on Economic Education.
- Walstad, W. B. & Rebeck, K. (2008). The Test of Understanding of College Economics. *American Economic Review*, 98, 547–551.
- Walstad, W. B., Rebeck, K. & Butters, R. B. (2013). *Test of economic literacy: Examiner's manual* (4th ed.). New York: Council for Economic Education.
- Walstad, W. B., Schmidt, S., Zlatkin-Troitschanskaia, O. & Happ, R. (2018). Pretest-posttest measurement of economic knowledge of undergraduates – Estimating guessing effects. Paper presented at the AEA Annual Meeting, Philadelphia, USA.
- Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Ashland, OH, US: Hogrefe & Huber Publishers.
- Wuttke, E., & Beck, K. (2002). Eingangsbedingungen von Studienanfängern – Die Prognostische Validität wirtschaftskundlichen Wissens für das Vordiplom bei Studierenden der Wirtschaftswissenschaften [Entrance conditions for first-year students – Prognostic validity of economic knowledge for the pre-diploma of students of economic sciences]. In K. Beck & K. Breuer, Arbeitspapiere WP [Working Papers Business Education], JGU Mainz. Accessed: https://download.uni-mainz.de/fb03-wipaed/ArbeitspapiereWP/gr_Nr.41.pdf (19th June 2019).

- Zlatkin-Troitschanskaia, O., Förster, M., & Kuhn, C. (2013). Modeling and measurement of university students' subject-specific competencies in the domain of business and economics – The ILLEV project. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education* (pp. 159–170). Rotterdam: Sense.
- Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Ed.), *Higher Education Learning Outcomes Assessment – International perspectives* (pp. 175–197). Frankfurt am Main: Lang.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Lautenbach, C., Molerov, D., Toepper, M., & Brückner, S. (2017). *Modeling and measuring competencies in higher education – Approaches to challenges in higher education policy and practice*. Wiesbaden: Springer.
- Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., Förster, M., & Brückner, S. (2019a). Validating a Test for Measuring Knowledge and Understanding of Economics Among University Students. *Zeitschrift für Pädagogische Psychologie*, 32 (2), 119-133.
- Zlatkin-Troitschanskaia, O., Schlax, J., Jitomirski, J., Happ, R., Kühling-Thees, C. & Pant, H. A. (2019b). Ethics and Fairness in Assessing Learning Outcomes in Higher Education. *Journal Higher Education Policy*, 32 (4), 537-556.

Appendix

A1 MLM with grade in business courses at T2 as dependent variable (group variable: 22 universities)

Variable	Model: Wald $\chi^2 = 199.70$, $p < 0.001$			
	b	SE	z	p
Constant	2.560	0.598	4.28	<0.001*
Attended courses in business	-0.434	0.091	-4.77	<0.001*
Attended courses in economics	0.015	0.055	0.28	0.782
Attended courses in microeconomics	-0.049	0.066	-0.75	0.455
Attended courses in macroeconomics	-0.040	0.074	-0.54	0.592
UEQ grade	0.586	0.070	8.32	<0.001*
Score BEFKI	-0.015	0.014	-1.07	0.285
Knowledge test score at T1	-0.036	0.010	-3.67	<0.001*
UEQ at school with B&E focus, no	-0.159	0.091	-1.75	0.080
B&E related vocational training, no	0.291	0.105	2.76	0.006*
Advanced course in B&E, no	0.039	0.074	0.53	0.598

Variable	Model: Wald $\chi^2 = 199.70, p < 0.001$			
	b	SE	z	p
Interest in B&E related topics	0.049	0.054	0.91	0.361
Gender, male	-0.057	0.072	-0.79	0.432
Age	-0.011	0.017	-0.65	0.513
Migration background, none	-0.123	0.081	-1.53	0.126
Preferred communication language, German	-0.041	0.264	-0.16	0.876

Note. $N = 529$ (due to missing values), UEQ = university entrance qualification, BE-FKI = fluid intelligence short scale, lower grades in Germany indicate better performance, * indicates significance on a 5 %-level.

A2 MLM with grade in economics courses at T2 as dependent variable (group variable: 22 universities)

Variable	Model: Wald $\chi^2 = 100.14, p < 0.001$			
	b	SE	z	p
Constant	1.611	0.848	1.90	0.057
Attended courses in business	-0.208	0.111	-1.87	0.061
Attended courses in economics	-0.203	0.116	-1.75	0.081
Attended courses in microeconomics	-0.126	0.097	-1.29	0.196
Attended courses in macroeconomics	0.006	0.101	0.06	0.953
UEQ grade	0.523	0.103	5.07	<0.001*
Score BEFKI	-0.007	0.020	-0.33	0.739
Knowledge test score at T1	-0.049	0.014	-3.47	0.001*
UEQ at school with B&E focus, no	-0.115	0.135	-0.85	0.394
B&E related vocational training, no	0.173	0.152	1.14	0.253
Advanced course in B&E, no	0.080	0.111	0.73	0.468
Interest in B&E related topics	0.049	0.081	0.61	0.541
Gender, male	0.055	0.104	0.53	0.594
Age	0.025	0.027	0.94	0.348
Migration background, none	-0.162	0.121	-1.34	0.180
Preferred communication language, German	0.831	0.373	2.23	0.026*

Note. $N = 387$ (due to missing values), UEQ = university entrance qualification, BE-FKI = fluid intelligence short scale, lower grades in Germany indicate better performance, * indicates significance on a 5 %-level.

A3 MLM with grade in microeconomics courses at T2 as dependent variable (group variable: 20 universities)

Variable	Model: Wald $\chi^2 = 76.74, p < 0.001$			
	b	SE	z	p
Constant	2.484	1.009	2.46	0.014*
Attended courses in business	-0.029	0.124	-0.23	0.815
Attended courses in economics	-0.046	0.091	-0.50	0.614
Attended courses in microeconomics	-0.375	0.164	-2.29	0.022*
Attended courses in macroeconomics	-0.161	0.119	-1.35	0.178
UEQ grade	0.592	0.120	4.94	<0.001*
Score BEFKI	-0.021	0.228	-0.90	0.368
Knowledge test score at T1	-0.041	0.016	-2.51	0.012
UEQ at school with B&E focus, no	-0.106	0.157	-0.67	0.501
B&E related vocational training, no	0.034	0.184	0.18	0.854
Advanced course in B&E, no	-0.213	0.123	-1.73	0.084
Interest in B&E related topics	0.031	0.091	0.34	0.733
Gender, male	0.022	0.120	0.18	0.855
Age	-0.000	0.032	-0.00	0.996
Migration background, none	-0.210	0.137	-1.54	0.124
Preferred communication language, German	0.613	0.387	1.58	0.113

Note. $N = 397$ (due to missing values), UEQ = university entrance qualification, BE-FKI = fluid intelligence short scale, lower grades in Germany indicate better performance, * indicates significance on a 5 %-level.

A4 Fit indices of grade-related multi-level models

Fit Indices	M0	p	M1	p	M2	p	M3	p	M4 (full)
Grade in business in T2									
	(N = 672)		(N = 534)		(N = 532)		(N = 529)		(N = 529)
Deviance	1,713.56	***	1,310.12	**	1,213.23	**	1,178.41	0.361	1,177.58
AIC	1,727.56		1,332.12		1,239.23		1,212.42		1,213.58
BIC	1,759.14		1,379.2		1,294.82		1,285.02		1,290.46
Grade in economics in T2									
	(N = 449)		(N = 393)		(N = 391)		(N = 387)		(N = 387)
Deviance	1,271.14	***	1,101.54	**	1,057.07	**	1,029.29	0.541	1,028.91
AIC	1,285.14		1,123.54		1,083.07		1,063.29		1,064.91
BIC	1,313.89		1,167.25		1,134.66		1,130.58		1,136.16
Grade in microeconomics in T2									
	(N = 491)		(N = 402)		(N = 400)		(N = 397)		(N = 397)
Deviance	1,510.39	0.312	1,233.70	**	1,189.81	*	1,173.22	0.733	1,173.11
AIC	1,524.39		1,255.70		1,215.81		1,207.22		1,209.11
BIC	1,553.77		1,299.66		1,267.70		1,274.95		1,280.82

Note. Deviance = 2*loglikelihood-value. * = sig. 5 %-level, ** = sig. on 1 %-level. M0 = control variables, M1 = attended courses added, M2 = intellectual ability added, M3 = learning opportunities added, M4 = interest added.



4.5

Influences on Master's Degree Students' Economic Knowledge

Kraitzek, A., Förster, M., and Zlatkin-Troitschanskaia, O.

Abstract

Despite the growing body of research, little is known about students' economic knowledge at the beginning of, or at various points throughout, their master's degree programs. In this study different impact factors on students' microeconomic knowledge are tested. Based on the "utilization of learning opportunities model" by Helmke we focus on 1) individual and sociodemographic characteristics of the students, 2) their learning potential and study-related characteristics and 3) the characteristics and structure of their learning environment. For each of the three influencing factors we used different indicators that were expected to correlate with students' knowledge. We tested the assumptions with hierarchical linear modeling with a sample of 1,281 students from 40 universities responding to microeconomics items. In the final model gender was the only significant variable among sociodemographic characteristics while study-related grades and attended courses were better predictors than the grade of the university entrance qualification.

Keywords

Higher education, economic knowledge, individual and learning environment factors, hierarchical linear modeling

Acknowledgements and Funding Information

The study was funded by the German Federal Ministry of Education and Research with the funding number 01PK11013. We would like to thank the anonymous reviewers for their constructive and valuable feedback.

1 Introduction

In higher education in Germany, a general tendency toward academization can be observed. Over the past several years, the number of first-semester bachelor students has grown continuously, and therefore also the number of graduates enrolling in master's degree study programs directly upon completing their bachelor studies (consecutive entrants) and graduates who have worked in the private sector and now seek professional development (lateral entrants) (e.g., Dionisius and Illger 2015). This is particularly the case in the study domain of business and economics (e.g., Zlatkin-Troitschanskaia et al. 2014, p. 175). The growing participation in these study programs has also led to greater diversity in student bodies at higher education institutes (HEIs). These developments in turn have led to greater expectations of higher education teachers, and to changes in learning opportunities.

Despite the growing body of research, little is known about students' economic knowledge at the beginning of, or at various points throughout, their master's degree programs (Happ et al. 2019). Imparting and transferring economic knowledge is a vital function of HEIs, and knowledge itself is paramount when launching a career (e.g., DIHK 2015). Deeper understanding of higher education students' development of knowledge is required to identify factors that influence learning (Seeber et al. 2015; Shavelson et al. 2018). Evidence on students' knowledge change would help determine implications for the organization of effective teaching and learning opportunities in higher education economics courses and programs (e.g., Siegfried and Wuttke 2016).

In this paper, influence factors on students' microeconomic knowledge are examined. First, a brief overview of the state of research is given and the theoretical-conceptual background of *knowledge* and how it is assessed in our study is described. Next, a description of the study and a multilevel analysis (MLA) is

provided. Finally, results of the study are discussed and implications for practical use and further research are explored.

2 Theoretical Background and Current State of Research

In previous research, various factors or combinations thereof leading to unfavorable preconditions have been identified that may influence the development of economic knowledge. According to the framework of the utilization of learning opportunities model of education by Helmke (2009, Figure 1), students' learning and perception of learning opportunities are influenced by various factors such as **(1)** the students' family and socialization environment as well as sociodemographic characteristics, **(2)** the students' learning potential and preconditions (e.g., prior knowledge, intelligence), and **(3)** structural factors (context) of the learning environment (e.g., type of HEI, such as university, university of applied science, or technical college, or the chosen course of study) in which learning takes place (Helmke and Schrader 2011). In this model, instruction, regardless of the field of study, is considered as merely an opportunity for students to learn something. This learning opportunity leads to actual learning, and how learning occurs depends on how the learning opportunity is perceived and interpreted by the student. Eventually, learning outcomes can be observed as a result of the complex interplay among the aforementioned factors. We focused on these three factor groups (1: Individual and Sociodemographic Characteristics, 2: Learning Potential and Study-related Characteristics, 3: Characteristics and Structure of the Learning Environment) in our study although learning activities and teacher- and instruction-related characteristics also influence students' development of knowledge. The goal of the paper is to investigate how sociodemographic characteristics, students' learning potential and preconditions as well as structural factors of the learning environment have an impact on the students' economic knowledge. We use different indicators for the three influencing factors that are expected to correlate with the dependent variable. These indicators and their expected impact are presented in Sections 2.2 to 2.4, after discussing the depending variable.

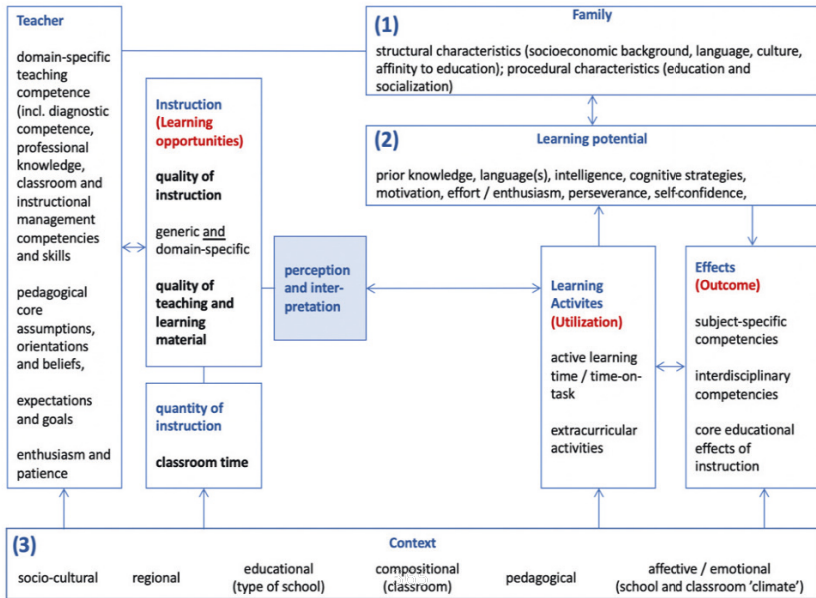


Figure 1 The utilization of learning opportunities model (Helmke 2009, p. 73)

2.1 Conceptual Fundamentals of Economic Knowledge

When it comes to modeling the latent construct of *economic knowledge*, a well-fund-ed theoretical approach, which is at the same time practically oriented, is essential. Individuals who possess economic knowledge should be able to handle situation that involve general economics in some way. Therefore, we understand economic knowledge as a domain-specific cognitive disposition which allows an individual to solve tasks that cover core content in a ‘typical’ principles course in economics (Walstad and Rebeck 2008, p. 549; Zlatkin-Troitschanskaia et al. 2015). With respect to both the cognitive requirements and the content areas of economics, the presumed factorial structure has been specified to model and to assess economic knowledge. The cognitive structure of economic knowledge was modeled by three cognitive levels: (1) recognition and understanding, (2) explicit application, and (3) implicit application. This approach, which is also in line with the model by Walstad et al. (2007, p.6), is based on an adapted version of Bloom’s taxonomy (Anderson and Krathwohl 2001). Concerning the level of content, the domain of economics

can be divided into the two sub-domains of macro- and microeconomics. Focusing on the latter, the content can be further categorized into “Basic Problems”, “Markets and Prices”, “Theory of Firm”, “Factor Markets”, the microeconomic “Role of Government” and the micro-dimension of “International exchange”. By this approach, following the model of Walstad, Watts and Rebeck (2007), the core curriculum in economics should be represented adequately (Zlatkin-Troitschanskaia et al. 2014). Microeconomic knowledge (i.e. general knowledge about the aforementioned categories) is thus considered the core component necessary to develop domain-specific (i.e., microeconomic) competences (Eberle et al. 2016; Siegfried and Wuttke 2016; Zlatkin-Troitschanskaia et al. 2014) and is essentially a result of multiple processes that are influenced by personal, social, and structural factors (Helmke and Schrader 2011; Sembill and Kärner 2018, p. 178).

2.2 Individual and Sociodemographic Characteristics

In previous research, various sociodemographic, and individual characteristics that may influence students' acquisition of economic knowledge at the master's degree level have been found. At an individual level, **gender** is considered one of the main predictors of developing economic knowledge. Numerous studies and large-scale assessments indicate that male students perform significantly better than their female counterparts in economics and economics-related subjects (e.g., Schumann and Eberle 2014; Förster et al. 2018a; Siegfried 2019; Zlatkin-Troitschanskaia et al. 2019a). In a comparison study of undergraduate students' performance on tests of micro- and macroeconomic content knowledge in Germany, Japan, and the United States, Brückner et al. (2015) found that male students outperformed female students (p. 510). In numerous other studies advantages for male test takers on economic knowledge tests were revealed (e.g., Asarta et al. 2014). As most studies in the field of economics have focused on bachelor students, little is known about the development of economic knowledge at the master's degree level. Happ et al. (2019) demonstrated that in the area of microeconomics, male students achieve significantly higher test scores than female students at the master's degree level as well.¹ Consequently, the first hypothesis can be formulated as follows:

1 For more information on research on the gender effect, see the special issue by Happ, Förster, and Siegfried (in review) where various hypotheses for the gender effect are tested.

H1a: Male students perform better in a microeconomic knowledge test than female students.

Due to the large influx of immigrants to Germany over the past several years, we do not know much about how these groups impact the diversity of the student body and its learning conditions. The socio-cultural and migration background and its implication for learning still poses a desideratum in higher education research and needs to be considered more intensely (Klaus 2020). Due to the current situation in Germany, the reason for migration and subsequently the term *learners' migration background* is often and mistakenly just referred to as flight migration, i.e. migration due to war or persecution in an individual's home country (Walwei 2016; Brücker et al. 2016). This narrow perspective on a learner's migration background excludes learners who visit German HEIs, for example, as part of student exchange programs or international scholarships. Hence, the term *migration background* has to be thought of on a broader basis, since it accumulates the increasing socio-cultural diversity of the student body. This is especially true for internationally oriented study subjects like economic and its related sub-areas (DAAD 2016, p. 15 and 18). However, studies conducted in Germany indicate that migration background as well as a home language other than German are unfavorable preconditions for knowledge acquisition and learning outcomes (e.g., Hurrelmann 2009; Happ et al. 2019). Home language acts as another predictor of economic knowledge (e.g., Brückner et al. 2015). Happ et al. (2019) identified negative effects of a migration background on students' economics knowledge acquisition in master's degree studies even after completion of a bachelor degree program.

H1b: Students without a migration background perform better in a microeconomic knowledge test than students with a migration background.

H1c: Students with German as the language spoken at home perform better in a microeconomic knowledge test than students with any other language than German spoken at home.

Educational choices also have an impact on students' learning outcomes, as does the amount of knowledge in economics they acquire prior to beginning higher education studies. As learning opportunities at secondary and vocational schools prior to higher education vary across the federal states of Germany, students acquire different knowledge in economics (e.g., Brückner et al. 2015). Thus, university entrance qualification grades are less comparable, and there is evidence that domain-specific preparatory training at commercial vocational schools specializing in business and economics influences the subsequent acquisition of economic knowledge significantly (e.g., Happ et al. 2017, p. 61ff.; Riebenbauer 2015a, p. 43f.).

H1d: Students who were trained at vocational schools with a special focus on business and economics perform better in a microeconomic knowledge tests than students from schools for general education.

After graduating from school, it is not uncommon for many young people in Germany to gain work experience before pursuing studies at college or university. Obtaining commercial vocational training and work experience in a field influences the acquisition of knowledge in that field. This has been proven in research on external accounting in which knowledge of double-entry bookkeeping and procurement and distribution processes were assessed (Fritsch et al. 2014, p. 37; Riebenbauer 2015b). Similar results were yielded in studies of internal accounting (Förster et al. 2016), finance (Förster et al. 2015), and micro- and macroeconomics (Zlatkin-Troitschanskaia et al. 2015). Although vocational training and years of work experience have a positive effect on students' performance on economic knowledge tests at the bachelor degree level (e.g., Happ et al. 2017), little is known about such effects at the master's degree level.

H1e: Students who completed a commercial vocational training perform better in a microeconomic knowledge test than students without commercial vocational training.

2.3 Learning Potential and Study-Related Characteristics

Research indicates that students entering a master's degree study program at university usually have completed previous studies and/or obtained special training and/or work experience and therefore have diverse knowledge (e.g., Happ et al. 2019). Particularly in the study domain of economics, (in addition to general skills), the domain-specific competences master's degree students gain from their prior education have significant effects on their acquisition of economic knowledge. Despite discussion about the explanatory power of grades (e.g., Trautwein et al. 2008), the university entry qualification grade is considered the most significant predictor of a person's learning potential and their anticipated academic performance before entering university (e.g., Steyer et al. 2005; Trapmann et al. 2007; Riebenbauer 2015b). The final grade at secondary school, for instance, has been found to influence bachelor students' scores on micro- and macroeconomic tests (e.g., Beck and Wuttke 2004; Schumann and Eberle 2014; Zlatkin-Troitschanskaia et al. 2015; Schlx et al. in this volume).

In addition to indicators of learning potential before starting a university study program, participation and grades in undergraduate studies are expected to be even stronger predictors of the development of economic knowledge during a mas-

ter's degree study program. We expect that the university entry qualification grade loses its predictive power the with increasing study progress. Results indicate that study-relevant performance like grades achieved in a bachelor degree program and performances in the relevant examinations could significantly explain the economics test performance of students in master's degree studies and are expected to be a better predictor than the university entry qualification grade (Troche et al. 2014). This could be justified with the Knowledge-Is-Power Hypothesis, which states that domain knowledge is the primary determinant of cognitive endeavors within a domain (Hambrick and Engle 2002).

H2a: The better the grade of a student's university entry qualification, the higher the student scores in a microeconomic knowledge test.

H2b: The better a student's final grade achieved in a bachelor degree program, the higher the student scores in a microeconomic knowledge test.

H2c: The explanatory power of the student's university entry qualification to explain microeconomic knowledge during the master's degree course is smaller than the final grade of the bachelor degree program.

Evidentially, learning opportunities in higher education are essential for the acquisition of domain-specific knowledge: In previous research, positive effects of (regularly) attending one or more lectures were found on students' acquisition of knowledge of external accounting (Fritsch et al. 2014), internal accounting (Förster et al. 2016), micro- and macroeconomics (Zlatkin-Troitschanskaia et al. 2015, p. 120 and p. 130; Schlax et al. in this volume; Siegfried 2019), and finance (Förster et al. 2015). While attending lectures is a variable depending on the specific structures at HEIs (Kühling-Thees et al. in this volume), performance on the examination at the end of a lecture is considered a qualitative indicator of the acquired knowledge and the students' learning performance (Shavelson et al. 2018).

H2d: Students who attended one or more lectures in microeconomics score higher in a microeconomic knowledge test than students who did not attend any lectures.

H2e: The better the grade a student receives on a microeconomic course, the better the student's performance in a microeconomic knowledge test.

2.4 Characteristics and Structure of the Learning Environment

According to the Helmke model, the characteristics and structure of the learning environment play a significant role with regard to students' learning activities. This has been especially evident since the introduction of the Bologna process,

which has required major reconstruction of structure and curricula over the past several years to meet the demands of an increasingly competence-oriented higher education. In higher education economics, different types of HEIs offer different learning opportunities, and the type of HEI students attend impacts the knowledge they acquire. University students exhibit greater economic knowledge than students of universities of applied science or technical colleges (e.g., Happ et al. 2013, p. 80; Zlatkin-Troitschanskaia et al. 2015, p. 128f.).

Although business and economics is a very broad study domain and economic content is also gaining importance in related subjects such as industrial engineering and business information systems (e.g., DIHK 2015), little is known about how master's degree students' economic knowledge varies across these subjects. Therefore, in the study presented here, we investigated whether and how various attended courses in economics-related study programs influence the state of students' economic knowledge. We expect that economics students should achieve the best results.

H3a: Students at universities reach higher scores in a microeconomic knowledge test than students at universities of applied sciences.

H3b: Students of economics perform better in a microeconomic knowledge test than students of any other economics-related subject.

Presents the superordinate conceptual model with all of the facets described above and the assumed correlations.

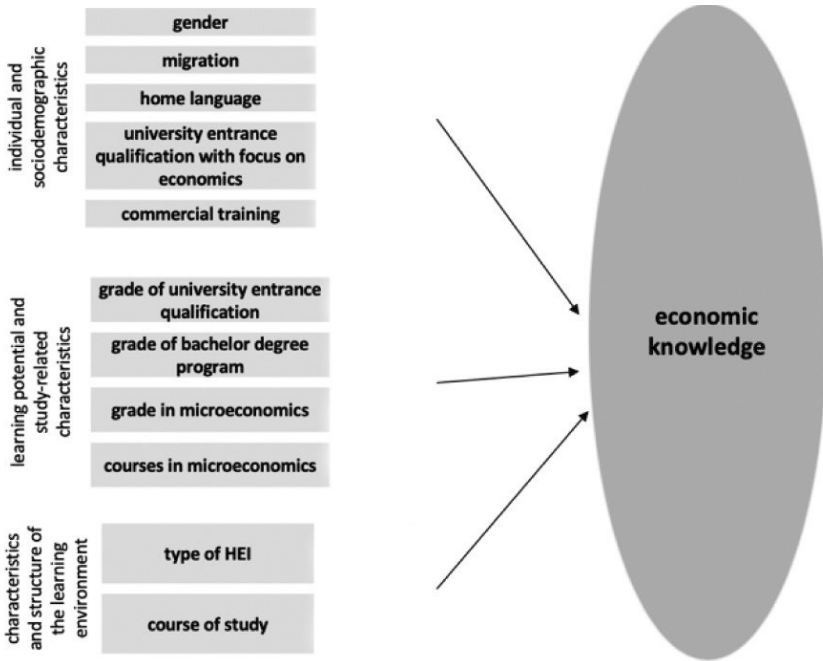


Figure 2 Conceptual Model

3 Study Design, Data Collection, and Analyses

3.1 Operationalization

To assess students' microeconomic knowledge, the comprehensively validated 4th version of the Test of Understanding College Economics (TUCE), which was originally developed by Walstad et al. (2007), was administered (Zlatkin-Troitschanskaia et al. 2015). In the WiWiKom project, the test was translated, validated by experts, and adapted for use in higher education in Germany (for the adaptation and validation processes, see Förster et al. 2014, Zlatkin-Troitschanskaia et al. 2014).

The microeconomics part of the TUCE consists of 30 items, each of which has one correct answer option of four (single choice). The items cover six subareas of microeconomics outlined by Walstad et al. (2007, p. 2: basic problems, markets and prices, theory of firm, factor markets, role of the government, and interna-

tional relations). In terms of cognitive demand and level of difficulty, the items are designed to assess recognition and understanding, explicit application, and implicit application and thus are seen as more or less equivalent to the three levels of Bloom's taxonomy, that is, remember, understand, and apply (Bloom et al. 1956, Krathwohl 2002, p. 216).

To counteract possible distortion effects (e.g., effects of students' fatigue during test taking, cognitive bias, etc.), a booklet design was chosen for data collection. In the study presented here, the microeconomics items along with finance and accounting items was spread over seven test booklets, ensuring an almost even distribution of microeconomics items in each booklet. The position of the part of the test concerning microeconomics was different in each booklet to minimize the possible effects of the positioning of items. The reliability of the TUCE had been proven for Germany and the US. For German higher education the EAP/PV-reliability for the microeconomics test was 0.717 in the booklet-design (for further information, see Zlatkin-Troitschanskaia et al. 2015). Furthermore, socio-demographic questions were included at the beginning of each booklet. With this procedure, an almost evenly split number of items in total across all booklets could be ensured. A brief overview of the complete booklet design is given in Table 1.

Table 1 Booklet design

Booklet No.:	1	2	3	4	5	6	7
Part 1	Socio	Socio	socio	Socio	socio	socio	socio
Part 2	finance 2	micro 1	micro 2	micro 3	account 1	account 2	finance 1
Part 3	micro 1	micro 2	micro 3	account 1	account 2	finance 1	finance 2
Part 4	micro 2	account 1	account 2	finance 1	finance 2	micro 1	micro 3
Total items	28	28	28	26	24	26	26

To identify and examine the factors influencing students' economic knowledge, in this study, students' microeconomic knowledge was operationalized through the share of overall number of microeconomics items they answered correctly to the number of microeconomics they had to answer in total (score_{mic}). The value of the score_{mic} differs between 0 (no items solved correctly) and 1 (all items solved correctly).

The independent variables were operationalized mostly via a single item in which test takers could choose a closed question requiring a *yes* or *no* answer (e.g., *Did you complete a vocational training program?*), a question requiring a single

response from a list (e.g., *What is your course of study?*), or an open-ended question requiring a short response. To determine home language test takers responded to one question about whether German was the language commonly spoken at home. Test takers were allowed to add some information about the language spoken primarily at home. Migration background was determined by the test takers' response to a question asking whether at least one parent did not originate from Germany. The average grade, for example, in a bachelor degree program, was supplied by students, who provided the numerical grade they received. Examples for various items of the independent variables and their response options are provided in Annex A.

3.2 Sample

In the WiWiKom project, data collection for the study presented here was conducted in summer 2015 as a one-off cross-sectional study at 40 HEIs in Germany. The method of the data collection, i.e. testing a variety of students who are enrolled in a variety of different HEIs across the country, indicates a hierarchical data structure which has to be taken into account in the data analysis (for Multi-Level Analysis, see Section 4.2). The booklets were administered to master's degree students in business and economics and related subareas, resulting in a sample of $N = 1,492$ students, with a total of $n = 1,281$ test takers having responded to the microeconomics items. This means that due to the booklet design, 85.9% of the test takers worked on the microeconomics part of the test. The average age of the participants was approximately 25 and the distribution across the genders was almost evenly split with $N = 678$ (i.e., 52.9%) male students. Of the 1,281 test takers, 342 (i.e., 26.7%) had a migration background and 491 (i.e., 38.3%) were attending a university of applied science. Other details about the sample are provided in Table 2.

Table 2 Sample description (test takers for microeconomics)

		M	SD	Min.	Max.	<i>n</i> (%)
Age		24.98	2.29	20	49	1255 (98.0)
Gender	Female					602 (47.0)
	Male					678 (52.9)
Migration background	Yes					342 (26.7)
	No					936 (73.1)
Home language	German					1095 (85.5)
	Other					181 (14.1)
UEQ with focus on economics	Yes					292 (22.8)
	No					985 (76.9)
Commercial training	Yes					252 (19.7)
	No					1026 (80.1)
UEQ grade*		2.24	.533	1.0	3.7	1252 (97.7)
Av. grade bachelor degree*		2.09	.430	1.0	3.5	1185 (92.5)
Av. grade microeconomics*		2.28	.737	1.0	4.0	938 (73.2)
Completed courses in microeconomics	None					208 (16.2)
	One					654 (51.1)
	More than one					333 (26.0)
Type of HEI	University					790 (61.7)
	University of applied science/technical college					491 (38.3)
Course of study	Economics and business administration					311 (24.3)
	Economics and business education					68 (5.3)
	Economics					137 (10.7)
	Business administration					577 (45.0)
	Business and industrial engineering					83 (6.5)
	Business information systems					48 (3.7)
	Other					53 (4.1)

Note. $N_{total} = 1,281$; 40 HEIs (23 univ., 16 univ. of appl. sc., 1 technical college).

*In the German grading system, grades range from 1 (best) to 6 (worst).

4 Results

4.1 Frequency of Correctly Answered Microeconomics Items

Of the $n = 1,281$ test takers who solved microeconomics items, only one was able to answer all of the questions correctly, thus scoring 100%. Overall, 175 test takers (i.e., 13.7%) scored 20% or less. In our sample, the average was $M = .437$ ($SD = .1828$), which means that 43.7% of all the questions about microeconomics were answered correctly. The overall distribution (classes) is described in Figure 3.

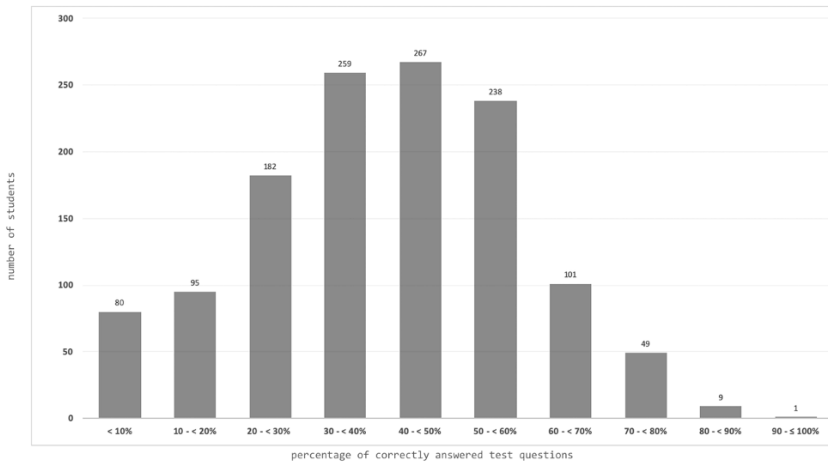


Figure 3 Overall distribution of correctly answered test questions

4.2 Results of the Multi-Level Analysis: Intra-Class Correlation

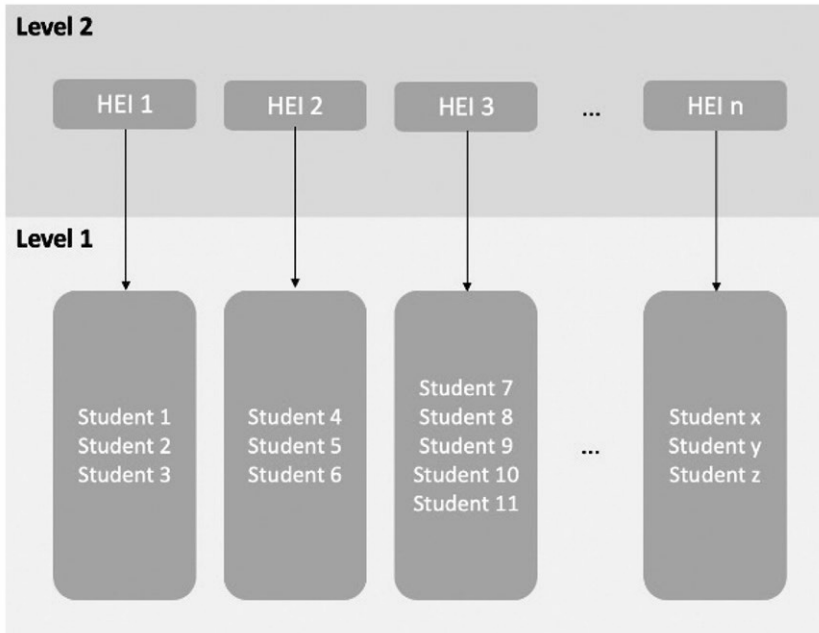


Figure 4 Hierarchical data structure

To test the hypotheses in this study, an MLA was conducted. As the students are allocated in different universities (Level 2), the hierarchical structure of the data was taken into consideration (Figure 4). In the present case, variance among the HEIs (Level 2) occurs as well as variance among the students of each HEI (Level 1). With an MLA being more complex than standard regression models, the first step prior to the main analysis was to investigate whether the hierarchical structure affected the results. To this end, we calculated a variance component model (Raudenbush and Bryk 2002) by which variance in microeconomic knowledge among the HEIs and the students is identified. The resulting intra-class correlation coefficient (ICC), which should be larger than 0.05 or 5% (Hox 2010; Heck et al. 2010), indicates whether conducting an MLA is justified. For our model, we computed the following variance component model and calculated an ICC of 11.75%, indicating that conducting an MLA is reasonable, and more than 11%

of the variance in the microeconomic test scores of the present sample can be described through differences among the HEIs alone. The calculations and results are displayed below.

$$\text{ICC: } \rho = \frac{\tau}{\tau + \sigma} = \frac{0,004048}{0,004018 + 0,030174} = 0.1175... \hat{=} 11.75\%$$

with $\tau = \text{Intercept (const.)}$ = variability across L2 (HEI)
 $\sigma = \text{Residual}$ = variability across L1 (students) (Baltes-Götz 2019).

4.3 Results

An overview of the results of the MLA is given in Table 3. We used a stepwise procedure to account for dependencies between the different groups of influencing factors; where in the first model we tested the individual and sociodemographic factors, in the second model we added learning potential and study-related characteristics of the students, and in the third model we added the structural factors of the learning environment.

In Model 1, which depicts individual and sociodemographic factors, the variables *gender* and *home language* correlated significantly with the microeconomic test scores. On average, male students solved approximately 6.7% more items correctly than female students, and students speaking a language other than German at home solved approximately 5% fewer items correctly than students speaking German at home, confirming H1a and H1c. *University entrance qualification from a school specializing in business and economics (H1d)* and *a completion of commercial training (H1e)* did not have a significant influence on the test scores. The variable *migration (H1c)* was almost significant.

In this model, 6.9% of the variance at the student level and 2.5% of the variance at the institutional level (L2: $R^2=.025$) could be explained by those variables. The rather small amount of variance could be explained that in this model only individual factors, that is Level-1 variables, were integrated.

In Model 2, when taking learning potential and study-related characteristics into account, the *University entrance qualification grade (H2a)*, *final grade in the bachelor degree study program (H2b)*, and *completion of one or more microeconomic courses (H2d)* had a highly significant effect on the students' test scores while the significant *gender (H1a)* effect found in Model 1 remained. The explanatory effect of the *University entrance qualification grade is lower than the effect of the final bachelor grade, which confirms H2c*.

Remarkably, the *grade in microeconomics* did not influence the test scores and was only close to being significant (H2e). Additionally, the effect of *home language*, which could be found in Model 1, no longer occurred when taking study-related characteristics into consideration (H1c). The effect of *migration* also was no longer observable. In this model, 14.4% of the variance could be explained by Level 1 variables (L1: $R^2 = .144$) and 49.1% by Level 2 variables (L2: $R^2 = .491$). Although we did not integrate any second-level variable until this point, the explanation of variance on the second level is higher than expected. This effect is due to the integration of the attended courses in microeconomics. Since students of universities of applied sciences have participated in less microeconomics courses, they received a microeconomics grade less often.

In Model 3, after adding structural factors of the learning environment, the significant gender effect (H1a) remained while no other sociodemographic or individual factors yielded substantial effects on the test scores (H1b-H1e). *Migration* (H1b) was close to being significant, as was the *UEQ grade* (H2a). However, the *final grade in the bachelor degree study program*, the *grade in microeconomics*, and *completion of courses in microeconomics* still showed considerable effects on microeconomic test scores.

In our sample, the *type of HEI*, which is the only pure second-level variable in our study, was only close to being significant and did not have as strong an effect as presumed in H3a. Considering the *course of study*, students of economics have the highest score and have a significant advantage compared to test takers from other fields of study. Only in comparison to economics and business education students ($p = .056$) and to business information systems students ($p = .061$) are the differences not significant on a 5%-Level. In this model, 15.6% of the variance could be explained by Level 1 variables (L1: $R^2 = .156$) and 79.3% by Level 2 variables (L2: $R^2 = .793$). The significant increase in variance explained by L2 variables here is due to the addition of the variable *type of HEI* in this model.

A general overview of the results embedded in the previous model of hypotheses is presented in Figure 5.

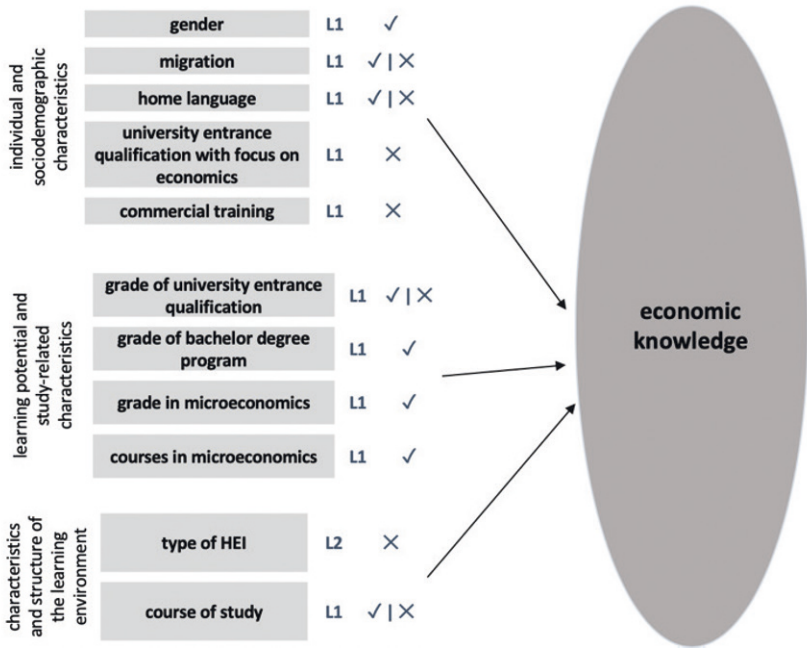


Figure 5 Results embedded in the conceptual model

Table 3 Results of the MLA

Variables	Base model	Model 1	Model 2	Model 3
fixed effects (SE)	.4409 (.0119)**	.3915 (.0208)***	.4639 (.0497)***	.4977 (.0608)***
Individual and sociodemographic characteristics				
gender (0 = female; 1 = male)		.0674 (.0099)***	.0578 (.0117)***	.0566 (.0117)***
migration background (0 = yes; 1 = no)		.0245 (.0130)†	.0204 (.0148)	.0256 (.0147)†
home language (0 = German; 1 = other)		-.0499 (.0167)**	-.0181 (.0203)	-.0128 (.0201)
UEQ with focus on EBA2 (0 = yes; 1 = no)		.0177 (.0120)	.0042 (.0143)	-.0000 (.0142)
completion of commercial training (0 = yes; 1 = no)		-.0131 (.0130)	-.0242 (.0152)	-.0223 (.0153)
Learning potential and study-related characteristics				
grade of UEQ3*				
grade of bachelor degree program*			-.0268 (.0120)*	-.0215 (.0120)†
grade in microeconomics*			-.0494 (.0165)**	-.0592 (.0164)***
courses in microeconomics attended (1 = attended)			-.0168 (.0090)†	-.0176 (.0089)*
Characteristics and structure of the learning environment			.0714 (.0128)***	.0588 (.0130)***
type of HEI (0=univ., 1=univ. of appl. sc./TC3)				-.0347 (.0189)†
course of study (reference: economics)				
economics and business administration				-.0930*** (.0211)
economics and business education				-.0631† (.0330)
business administration				-.0931*** (.0200)
business and industrial engineering				-.1255** (.0419)
business information systems				.0876† (.0467)
"Others"				-.0959* (.0393)
random effects (SE)/expl. of variance				
variance (L2)/pseudo R2 (%)	.0040 (.001)***/--	.0039 (.001)***2.5	.0020 (.001)*49.1	.0008 (.001)79.3
variance (L1)/pseudo R2 (%)	.0302 (.001)***/--	.0281 (.001)***/6.9	.0258 (.001)***/14.4	.0255 (.001)***/15.6

Note. $p \leq .10$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$. (Values in brackets represent standardized errors); 2EBA = economics and business administration; 3TC=technical college; *In the German grading system, grades range from 1 (best) to 5 or 6 (worst).

5 Conclusion

5.1 Discussion

The three examined groups of major influences impacted master's degree students' economic knowledge to a different extent. Within the group of *sociodemographic and individual influences*, a significant gender effect was found, which influenced performance on the microeconomic knowledge test significantly. This effect was observable at the beginning of a master's degree study program, as male learners performed significantly better on a standardized test of microeconomic knowledge. Although this finding is in line with results of prior studies of bachelor's and master's degree students (e.g., Zlatkin-Troitschanskaia et al. 2019a; Happ et al. 2019), the question arises as to why gender still had an effect at the master's degree level. One reason might be that microeconomic knowledge is heavily structured by implicit rules of logic and based on mathematical content, which might be favored by male learners (Ballard and Johnson 2005; Shavelson et al. 2019). Further, it could be argued that interest plays a large role in the acquisition of knowledge. Therefore, if male learners are more interested in this field than their female peers, they might be better at acquiring microeconomic knowledge (Förster et al. 2018). Another possible explanation, as the results of prior studies suggest, might lie in the test format, which is a simple 1-out-of-4 multiple (single) choice and may be more suitable for males (e.g., Walstad and Robson 1997; Arnold and Rowaan 2014; Ackermann and Siegfried in review). One way to test the latter assumption would be to redesign the test to a format that requires more elaborate, complex answers to the same test items than simple right-or-wrong answers. In this context, it might be worthwhile to investigate if the level of difficulty of the items can be varied in terms of test format, for example, how students perform when the test items contain predominantly numbers (i.e., arithmetical forms) or are purely text-based (Brückner et al. 2015; Förster et al. 2015). Against the background of greater competence-oriented education in Germany, it also might be necessary in the future to develop competence- and performance-oriented achievement tests designed to assess not only microeconomic knowledge but knowledge in other subjects or competence on a broader basis.

Migration background and *home language* had an impact on economic knowledge when only sociodemographic variables were included in the model. Students who had at least one parent not born in Germany or who spoke a language other than German at home performed worse than students without a migration background. These differences became smaller and were no longer significant when other factors were controlled. This might indicate that especially the final grade in

a bachelor degree study program, the completion of a microeconomic course, and the HEI institution as well as the course chosen might mediate some of the effects of migration background (Happ et al. 2019).

Concerning *prior economic education*, completion of commercial (vocational) training did not correlate with microeconomic knowledge in the master's degree study program. Since the subarea of microeconomics is, in contrast to fundamental economics, not extensively addressed in most commercial and vocational traineeships (Förster et al. 2018a), this finding is meaningful. However, as prior commercial training generally plays a significant role in the acquisition of knowledge in business and economics, it should be examined whether any effects can be observed in business-related areas such as accounting, finance, and controlling (Happ et al. 2019).

In terms of *cognitive conditions* as an important part of *study-related characteristics*, the effect of students' UEQ grade was no longer significant when their final grade in their bachelor degree study program and their grade in a microeconomics course were taken into account. This supports the assumptions that the final grade at secondary school becomes less closely linked to performance over the course of studies and that study-related grades are better predictors of students' knowledge. The correlation between the final grade in a microeconomics course and test results and the correlation between attendance in a microeconomics course and test results underline the validity of the TUCE.

At the *structural level of HEIs*, results of the study presented here indicate that (regularly) attending lectures on microeconomics has a significant effect on the acquisition of microeconomic knowledge. This finding is in line with results of previous studies (e.g., Stanca 2006). However, as the current study also shows, master's degree students in Germany were able to answer only approximately 44 % of the test items correctly. Since in the study presented here more than 10 % of the variance in microeconomic knowledge could be explained by the type of HEI alone, the questions arise as to what extent the HEI itself as an institution matters and whether the actual process of knowledge transfer in HEIs (e.g., during lectures) should be reconsidered. Previous research revealed that students attending university have greater microeconomic knowledge than students attending universities of applied science (Zlatkin-Troitschanskaia et al. 2015, p. 130). This might partly be due to the fact that the different types of HEIs differ in learning opportunities. At universities of applied science focus usually is not solely on economics; therefore, there are fewer economics-related topics covered and less economics-related content learned. However, results of curricular analyses indicate that the overall amount of economic content taught at universities of applied science is generally lower than at universities (Zlatkin-Troitschanskaia et al. 2015, p. 130).

Furthermore, students of economics performed better on the microeconomic knowledge test than students of other subjects such as business, which indicates that the course of study is relevant. This is an indicator that the chosen test is valid as the economics students performed best due to their learning opportunities in economics compared to students of other degree courses.

5.2 Limitations, Implications, and Outlook

Internal differentiation at the HEI level, that is, adjustments to learning opportunities, are required due to the increasing heterogeneity of students bodies (Happ et al. 2017, p. 74f.). This is the blind spot of the model used and data gathered in this study, as details about instruction in the classroom, learning activities in and out of class, or even teacher characteristics were unknown; the assessed study-related characteristics such as grades and attended courses were self-reported.

Another limitation is the drawn sample as it was not representative, neither on the institutional nor on the individual level. Though the sample was substantial in terms of the number of participating universities and students, it cannot be guaranteed that the results can be generalized across German higher education. Moreover, the results presented in this paper focus on economics, and here on microeconomics only. As Happ et al. (2019) demonstrated, in other business and economics subdimensions results might be different.

Furthermore, this study has a one-term cross-sectional design. Drawing conclusions about how the acquisition of microeconomic knowledge is influenced by specific variables in the long run requires longitudinal research like in the WiWiKom II project (Zlatkin-Troitschanskaia et al. 2019b; Schlax et al. in this volume). In the WiWiSET project, the current data is provided combined with investigations into academic success and failure (e.g., drop-out rates) and reasons for them (Kühling-Thees, in this volume). In addition to large-scale assessments like in the WiWiKom and WiWiSET projects, conducting microstudies, that is, (more qualitative) investigations at the classroom or course level, might be an effective way to gain deeper understanding of how knowledge develops over time and which specific learning environment fosters the desired learning outcomes (for a pre-test eye-tracking study with economics students, see, e.g., Klein et al. 2019).

Due to test limitations and time restrictions, we had to focus on selected variables within the three influencing factors presented in this paper. Therefore, only some aspects of the Helmke model have been taken into account and we approached the model from a rather formal learning opportunities perspective. There are more indicators that may have a significant impact on students' economic knowledge (e.g. socioeconomic background, economic socialization, study

motivation, self-confidence) and those should be examined in future studies in a more differentiated way.

Despite these limitations, the results from the MLA indicate, in particular, a need to explore and discuss the quality of instruction in higher education, as conditions vary significantly among and within universities. The instruction- and learning-activities-level within the utilization of learning opportunities model has to be focused more closely in microstudies. We do not know how students used the formal learning opportunities and what happened within the classroom. However, both instruction methods and learning activities are closely related to the context within and between the universities, which should be examined in further details. While at many universities obligatory lectures can be attended by more than 300 students, in other very specialized courses the groups of learners are quite small. In addition, groups of students normally are smaller at universities of applied sciences and technical colleges than at universities. These conditions have a strong impact on teachers' choice of instructional method and the learning culture. Revamped test and course designs as well as a restructuring of feedback methods also might be essential. Formative feedback strategies, for example, continuous and ongoing electronic feedback, concerning the state of knowledge and learning outcomes might enhance students' acquisition of knowledge even in lectures attended by numerous students (Förster et al. 2018b). In addition, promoting language comprehension and communication skills might foster general understanding of economic topics such as technical terms and domain-specific expressions; such measures would be helpful especially for learners with a migration background. Additionally, in higher education, the teaching competence of the instructor and their pedagogical skills should have an important impact on the effect and the utilization of the offered learning opportunities. While there are no obligated trainings for lecturers in higher education, the variance of lecturers' teaching competences can be expected to be higher than in school and the impact on students' outcome has to be examined in further studies. The implementation of these measures would permit investigation into individual factors that influence knowledge, knowledge acquisition, and learning outcomes.

References

- Ackermann, N., & Siegfried, C. (2019). Die Bedeutung schulischer Lerngelegenheiten für die wirtschaftsbürgerliche Kompetenz: Ein Vergleich von Gymnasialschülerinnen und -schülern in der Schweiz und in Deutschland. [The importance of educational learning opportunities for economic competencies: a comparison between A-Level students in Switzerland and Germany]. *Zeitschrift für ökonomische Bildung*.
- Arnold, I. J. M., & Rowaan, W. (2014). First-Year Study Success in Economics and Econometrics: The Role of Gender, Motivation, and Math Skills. *Journal of Economic Education*, 45(1), 25–35.
- Asarta, C., Butters, R.B., & Thompson, E. (2014). The gender question in economic education: Is it the teacher or the test?. *Perspectives on Economic Education Research*, 9, 1–19.
- Ballard, C., & Johnson, M. (2005). Gender, Expectations, and Grades in Introductory Microeconomics at a U.S. University. *Feminist Economics*, 11(1), 95–122.
- Baltes-Götz, B. (2019). *Analyse von hierarchischen linearen Modellen mit SPSS. [Analysis of hierarchical linear models with SPSS]*. (Zentrum für Informations-, Medien- und Kommunikationstechnologie) Universität Trier. <https://www.uni-trier.de/fileadmin/urt/doku/hlm/hlm.pdf>. Accessed: 30.07.19.
- Beck, K., & Wuttke, E. (2004). Eingangsbedingungen von Studienanfängern: Die Prognostische Validität wirtschaftskundlichen Wissens für das Vordiplom bei Studierenden der Wirtschaftswissenschaften. [Input conditions of university entrants: prognostic validity of economic knowledge for the intermediate diploma of economic students]. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 100(1), 116–124.
- Bloom, B. S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., and Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay.
- Brücker, H., Rother, N., & Schupp, J. (2016). *IAB-BAMF-SOEP-Befragung von Geflüchteten: Überblick und erste Ergebnisse. Forschungsbereich 29. [IAB-BAMF-SOEF refugee survey: overview and first results. Research report 29]*. Nürnberg: Bundesamt für Migration und Flüchtlinge BAMF. https://www.bamf.de/SharedDocs/Anlagen/DE/Publikationen/Forschungsberichte/fb29-iab-bamf-soep-befragung-gefluechtete.pdf?__blob=publicationFile. Accessed: 13.09.2019.
- Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., Happ, R., Walstad, W. B., Yamaoka, M., & Asano, T. (2015b). Gender Effects in Assessment of Economic Knowledge and Understanding: Differences Among Undergraduate Business and Economics Students in Germany, Japan, and the United States. *Peabody Journal of Education*, 90(4), 503–518.
- DAAD – German Academic Exchange Service (2016). *Wissenschaft weltoffen. Daten und Fakten zur Internationalität von Studium und Forschung in Deutschland. [Facts and Figures on the International Nature of Studies and Research in Germany]*. Bielefeld: W.Bertelsmann Verlag. http://www.wissenschaftweltoffen.de/publikation/wiwe_2016_verlinkt.pdf. Accessed: 13.09.2019.
- Deutscher Industrie- und Handelskammertag DIHK (2015). *Kompetent und praxisnah – Erwartungen der Wirtschaft an Hochschulabsolventen. Ergebnisse einer DIHK Online-Unternehmensbefragung. [Skillful and experienced: Economy's expectations towards university graduates. Results of a survey by the Chamber of Industry and Com-*

- merce]. https://www.dihk.de/ressourcen/downloads/dihk-umfrage-hochschulabsolventen-2015.pdf/at_download/file?mdate=1453731575017. Accessed: 19.12.2017.
- Dionisius, R., & Illger, A. (2015). Mehr Anfänger/-innen im Studium als in Berufsausbildung. [More university beginners than in vocational training]. *Zeitschrift des Bundesinstituts für Berufsbildung*, BWP, 44(4), 43–46.
- Eberle, F., Schumann, S., Kaufmann, E., Jüttler, A., & Ackermann, N. (2016). Modellierung und Messung wirtschaftsbürgerlicher Kompetenz von kaufmännischen Auszubildenden in der Schweiz und in Deutschland (CoBALIT). [Modelling and assessment of economic competence of vocational trainees in Switzerland and Germany (CoBALIT)]. In K. Beck, M. Landenberger and F. Oser (Eds.), *Technologiebasierte Kompetenzmessung in der beruflichen Bildung*. Ergebnisse aus der BMBF-Förderinitiative ASCOT (pp. 93–117). Bielefeld: W. Bertelsmann Verlag.
- Förster, M., Zlatkin-Troitschanskaia, O., Brückner, S., & Hiber, J. (2014). *Adapting and Validating the Test of Understanding in College Economics to Assess the Economic Knowledge and Understanding of Students in Germany (Discussion Paper; Annual Meeting of the American Economic Association)*. Philadelphia: AEA. <http://www.aeaweb.org/aea/2014conference/program/retrieve.php?pdfid=312>.
- Förster, M., Brückner, S., & Zlatkin-Troitschanskaia, O. (2015). Assessing the financial knowledge of university students in Germany. *Empirical Research in Vocational Education and Training*, 7(6), 1–20.
- Förster, M., Brückner, S., Beck, K., Zlatkin-Troitschanskaia, O., & Happ, R. (2016). Individuelle und kontextuelle Prädiktoren des Fachwissenserwerbs zum Internen Rechnungswesen im Hochschulstudium. [Individual and contextual predictors for knowledge acquisition in the field of internal accounting in higher education]. *Zeitschrift für Erziehungswissenschaften*, 19(2), 375–393.
- Förster, M., Happ, R., & Maur, A. (2018a). The relationship among gender, interest in financial topics and understanding of personal finance. In M. Förster, R. Happ, W. B. Walstad, and C. J. Asarta, (Eds.), *Financial Literacy* (pp. 293–309). Landau: Verlag Empirische Pädagogik.
- Förster, M., Weiser, C., & Maur, A. (2018b). How Feedback Provided by Electronic Quizzes Affects Learning Outcomes of University Students in Large Classes. *Computers and Education*, 121, 100–114.
- Fritsch, S., Seifried, J., Wuttke, E., & Fortmüller, R. (2014). Zum Einfluss von Lerngelegenheiten auf Fachwissen und fachdidaktisches Wissen von angehenden Lehrern und Lehrerinnen – das Beispiel Wirtschaftspädagogik. [The influence of learning opportunities on content knowledge and pedagogical knowledge of future teachers – an example from the area of Business and Economic Education]. *wissenplus*, 5, 31–35.
- Hambrick, D., & Engle, R. (2002). Effects of Domain Knowledge, Working Memory Capacity, and Age on Cognitive Performance: An Investigation of the Knowledge-Is-Power Hypothesis. *Cognitive Psychology*, 44(4), 339–387.
- Happ, R., Nagel, M.-T., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2019). How migration background affects master degree students' knowledge of business and economics. *Studies in Higher Education*, 1–16 (online first).
- Happ, R., Schmidt, S., & Zlatkin-Troitschanskaia, O. (2013). Der Stand des wirtschaftswissenschaftlichen Fachwissens von Bachelorabsolventen der Universität und der Fachhochschule. [The status quo of economic knowledge of bachelor students in university

- and university of applied science]. In U. Faßhauer, B. Fürstenau and E. Wuttke (Eds.), *Jahrbuch der berufs- und wirtschaftspädagogischen Forschung* (pp. 73–85). Opladen: Barbara Budrich.
- Happ, R., Schmidt, S., Zlatkin-Troitschanskaia, O., & Förster, M. (2017). Einfluss des Migrationshintergrunds bei Studierenden auf den Fachwissenserwerb im wirtschaftswissenschaftlichen Studium – eine vergleichende Längsschnittanalyse. [Influence of students' migration background on knowledge acquisition during the study of Economics – a comparative longitudinal analysis]. *Zeitschrift für Bildungsforschung*, 7(1), 59–77.
- Heck, R. H., Thomas, S. L., & Tabata, L. N. (Eds.). (2010). *Multilevel and Longitudinal Modeling with IBM SPSS*. New York, London: Routledge Taylor and Francis Group.
- Helmke, A. (2009). Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts. [Educational quality and teacher's professionalism. Diagnostics, evaluation and improvement of education]. Seelze-Velber: Klett | Kallmeyer, Erhard Friedrich Verlag.
- Helmke, A., & Schrader, F-W. (2011). Vom Angebots-Nutzungs-Modell zur Unterrichtsentwicklung. [From the utilization-of-learning-opportunities-model towards the evolution of education]. In A. Bartz, H.-J. Brandes and S. Engelke (Eds.), *Praxishilfen für die mittlere Führungsebene in der Schule. Modul 3: Unterrichtsentwicklung* (pp. 3–6). Köln: Carl Link Verlag.
- Hox, J. J. (Eds.). (2010). *Quantitative methodology series. Multilevel analysis: Techniques and applications*. New York: Routledge/Taylor and Francis Group.
- Hurrelmann, K. (2009). Ökonomische Bildung an Schulen: Ein innovativer Ansatz zur Förderung auch der benachteiligten SchülerInnen. [Economic education at schools: an innovative approach to support handicapped students]. *ZWD-Magazin*, 11, 8–12.
- Klaus, S. (2020). *Biographische Konstruktionen zur Ambivalenz von Hochschulzugang und Fluchthintergrund. [Biographical constructions concerning the ambivalence between university access and flight]*. Wiesbaden: Springer VS.
- Klein, P., Küchemann, S., Brückner, S., Zlatkin-Troitschanskaia, O. & Kuhn, J. (2019). Student understanding of graph slope and area under a curve: A replication study comparing first-year physics and economics students. *Physical Review Physics Education Research*, 15, 020116.
- Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4), 212–218.
- Raudenbush, S. W., & Bryk, A.S. (Eds.). (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage.
- Riebenbauer, E. (2015a). Analyse von Einflussfaktoren auf das Rechnungswesenwissen von Studierenden der Wirtschaftspädagogik. [Analysis of influencing factors on knowledge in accounting of students of Business and Economic Education]. *Wissenplus*, 14/15(5).
- Riebenbauer, E. (2015b). *Lehr-Lern-Prozesse im Rechnungswesen – Forschungsdesign und erste Analysen zum Fachwissen von Studierenden der Wirtschaftspädagogik. [Teaching and learning processes in accounting – research designs and first analyses on content knowledge of students of Business and Economic Education]*. http://www.bwpat.de/ausgabe28/riebenbauer_bwpat28.pdf. Accessed: 22.02.2019.
- Schumann, S., & Eberle, F. (2014). Ökonomische Kompetenzen von Lernenden am Ende der Sekundarstufe II. [Learners' economic competencies by the end of upper secondary level]. *Zeitschrift für Erziehungswissenschaften*, 17(1), 103–126.

- Seeber, S., Schumann, S., & Nickolaus, R. (2015). Ökonomische Bildung: Konzeptuelle Grundlagen und empirische Befunde. [Economic education: conceptual basics and empirical results]. In G. Weißenö and C. Schelle (Eds.), *Empirische Forschung in gesellschaftswissenschaftlichen Fachdidaktiken* (pp. 169–183). Wiesbaden: Springer.
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Marino, J. (2018). Performance Indicators of Learning in Higher-Education Institutions: Overview of the Field. In E. Hazelkorn, H. Coates, & A. Cormick (Eds.), *Research Handbook on Quality, Performance and Accountability in Higher Education* (pp. 249–263). Cheltenham: Edward Elgar.
- Shavelson, R. J., Marino, J., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2019). Reflections on the Assessment of Quantitative Reasoning. In L. Tunstall, G. Karaali & V. Piercey (Eds.), *Shifting Contexts, Stable Core: Advancing Quantitative Literacy in Higher Education* (pp. 163–176) Washington, DC: Mathematical Association of America.
- Siegfried, C. (2019). Wirtschaftswissenschaftliche Lerngelegenheiten als notwendiger Bestandteil der universitären Ausbildung von allgemeinbildenden Lehramtsstudierenden in der Domäne Wirtschaft. [Economic learning opportunities as a necessary element in university education of future Economics teachers]. *Zeitschrift für Erziehungswissenschaften*, 22(3), 593–616.
- Siegfried, C., & Wuttke, E. (2016). How can Prospective Teachers Improve their Economic Competence?. *Zeitschrift für ökonomische Bildung*, 4, 65–86.
- Stanca, L. (2006). The Effects of Attendance on Academic Performance: Panel Data Evidence for Introductory Microeconomics. *Journal of Economic Education*, 37 (3), (pp. 251–266).
- Steyer, R., Yousfi, S., & Würfel, K. (2005). *Prädiktoren von Studienerfolg*. [Predictors of study success]. *Psychologische Rundschau*, 56(2), 129–131.
- Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse. [The validity of school grades as predictor of study success – a meta-analysis]. *Zeitschrift für pädagogische Psychologie*, 21 (1), 11–27.
- Trautwein, U., Lüdtke, O., Becker, M., Neumann, M., & Nagy, G. (2008). Die Sekundarstufe I im Spiegel der empirischen Bildungsforschung: Schulleistungsentwicklung, Kompetenzniveaus und die Aussagekraft von Schulnoten. *Ausbildungsfähigkeit im Spannungsfeld zwischen Wissenschaft, Politik und Praxis*, 91–107.
- Troche, S. J.; Mosimann, M., & Rammsayer, T. H. (2014). Die Vorhersage des Studienerfolgs im Masterstudiengang Psychologie durch Schul- und Bachelorstudienleistungen. [prediction of study success in master's degree course psychology with school grades and grades of the bachelor degree course]. *Beiträge zur Hochschulforschung*, 36(1), 30–45.
- Walstad, W. B., & Rebeck, K. (2008). The test of understanding of college economics. *American Economic Review: Papers & Proceedings*, 98(2), 547–551.
- Walstad, W. B., & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *Journal of Economic Education*, 28, (155–71).
- Walstad, W. B., Watts, M., & Rebeck, K. (Eds.). (2007). *Test of understanding college economics. Examiner's manual*. New York: National Council of Economic Education.
- Walwei, U. (2016): Flucht und Migration: Herausforderungen für Bildung, Ausbildung und Arbeitsmarktpolitik. [Flight and migration: challenges for education, training and labor policy]. *Arbeit*, 25(3–4), 169–194.

- Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Eds.), *Higher education learning outcomes assessment: international perspectives* (pp. 175–197). Frankfurt am Main: Peter Lang International.
- Zlatkin-Troitschanskaia, O., Förster, M., Schmidt, S., Brückner, S., & Beck, K. (2015). Erwerb wirtschaftswissenschaftlicher Fachkompetenz im Studium. Eine mehrbenenanalytische Betrachtung von hochschulischen und individuellen Einflussfaktoren. [Acquisition of economic competence during study. A multilevel-analytic view on academic and individual influence factors]. *Zeitschrift für Pädagogik*, 61, 166–135.
- Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., Pant, H.A., Förster, M., & Brückner, S. (2019a). Validating a Test for Measuring Knowledge and Understanding of Economics Among University Students. *Zeitschrift Pädagogische Psychologie*, 32(2), 119–133.
- Zlatkin-Troitschanskaia, O., Schlax, J., Jitomirski, J., Happ, R., Kühling-Thees, C., Brückner, S., & Pant, H. A. (2019b). Ethics and fairness in assessing learning outcomes in higher education. *Higher Education Policy*, 32, 537–556.

Annex

Table A Examples for operationalization of independent variables taken from the booklet

1. Gender

Female.....

Male

Item 1: Gender

3. Which language is most frequently spoken in your family environment?

German

Other: _____

Item 3: Home language

16. What is your final average bachelor grade? _____.____

Item 16: Average grade of bachelor degree

18. How many courses of the following subjects did you visit prior to the current semester (e.g. in your bachelor studies) and which grade did you receive in each respective first exam?

	none	one	two or more	average grade
Microeconomics and adjoining sub-areas, e.g. industrial economics, market prices and competition	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	-----
completed courses; average grade				



Correction to: Measuring Scientific Reasoning Competencies – Multiple Aspects of Validity

Krüger, D., Hartmann, S., Nordmeier, V., and Upmeier zu Belzen, A.

Correction to:
Chapter 13 in: O. Zlatkin-Troitschanskaia et al. (Hrsg.),
***Student Learning in German Higher Education*,**
https://doi.org/10.1007/978-3-658-27886-1_13

In the electronic edition of this chapter the wrong university name (Freie Universität Berlin) was indicated for the authors S. Hartmann and A. Upmeier zu Belzen. This has now been corrected to "Humboldt-Universität zu Berlin".

The updated version of this chapter can be found at
https://doi.org/10.1007/978-3-658-27886-1_13