

Methodology of Educational Measurement and
Assessment

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

Assessment of Learning Outcomes in Higher Education

Cross-National Comparisons and
Perspectives

 Springer

Methodology of Educational Measurement and Assessment

Series Editors

Bernard Veldkamp, Research Center for Examinations and Certification (RCEC),
University of Twente, Enschede, The Netherlands

Matthias von Davier, National Board of Medical Examiners (NBME), Philadelphia,
USA

This book series collates key contributions to a fast-developing field of education research. It is an international forum for theoretical and empirical studies exploring new and existing methods of collecting, analyzing, and reporting data from educational measurements and assessments. Covering a high-profile topic from multiple viewpoints, it aims to foster a broader understanding of fresh developments as innovative software tools and new concepts such as competency models and skills diagnosis continue to gain traction in educational institutions around the world. *Methodology of Educational Measurement and Assessment* offers readers reliable critical evaluations, reviews and comparisons of existing methodologies alongside authoritative analysis and commentary on new and emerging approaches. It will showcase empirical research on applications, examine issues such as reliability, validity, and comparability, and help keep readers up to speed on developments in statistical modeling approaches. The fully peer-reviewed publications in the series cover measurement and assessment at all levels of education and feature work by academics and education professionals from around the world. Providing an authoritative central clearing-house for research in a core sector in education, the series forms a major contribution to the international literature.

More information about this series at <http://www.springer.com/series/13206>

Olga Zlatkin-Troitschanskaia
Miriam Toepper • Hans Anand Pant
Corinna Lautenbach
Christiane Kuhn
Editors

Assessment of Learning Outcomes in Higher Education

Cross-National Comparisons
and Perspectives

 Springer

Editors

Olga Zlatkin-Troitschanskaia
Johannes Gutenberg University
Mainz, Germany

Miriam Toepper
Johannes Gutenberg University
Mainz, Germany

Hans Anand Pant
Humboldt-Universität zu Berlin
Berlin, Germany

Corinna Lautenbach
Humboldt-Universität zu Berlin
Berlin, Germany

Christiane Kuhn
Johannes Gutenberg University
Mainz, Germany

ISSN 2367-170X ISSN 2367-1718 (electronic)
Methodology of Educational Measurement and Assessment
ISBN 978-3-319-74337-0 ISBN 978-3-319-74338-7 (eBook)
<https://doi.org/10.1007/978-3-319-74338-7>

Library of Congress Control Number: 2018930984

© Springer International Publishing AG, part of Springer Nature 2018

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, express 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.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Introduction: Assessing Student Learning Outcomes in Higher Education

Until the last decade, objective information on student learning and student learning outcomes in higher education at the national and international levels was scarce. This area was largely underrepresented in comparison to other areas of formal education such as school. In the context of current developments in higher education such as internationalization of study programs and ever-increasing student mobility and the ensuing increase in heterogeneity of students' learning conditions, the need for objective, valid, and reliable assessment tools that adhere to the Standards for Educational and Psychological Testing set out by the American Educational Research Association (AERA) has become urgent. This has led to intense research efforts being made within and across many countries, which are of great practical and political importance.

This book presents the most significant of these initiatives and developments in order to highlight the tremendous work national and international research communities have done in this area over the past decade. A broad range of national and international assessment research projects and curricular innovation initiatives in higher education focusing on both domain-specific and generic student learning outcomes are presented in this volume. Results and lessons learned from various research programs such as the German Modeling and Measuring Competencies in Higher Education (KoKoHs) and feasibility studies such as the Assessment of Higher Education Learning Outcomes (AHELO; an international comparative study by the Organisation for Economic Co-operation and Development (OECD) of students' generic skills and economic and engineering competencies) form the basis of several ongoing initiatives by testing institutes to make assessments suitable for use in higher education abroad. Examples include the Educational Testing Service's (ETS) Heighten Outcomes assessment and the Council for Aid to Education's (CAE) Collegiate Learning Assessment CLA+. At the European level, the CALOHEE initiative on Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe aims to develop a joint basis for learning outcomes in higher education as well as curricula in five disciplines, including education. One of the most current international initiatives, the International Collaborative for Performance Assessment of Learning in Higher Education (iPAL),

focuses on developing performance assessments of learning that meet high standards in psychometric quality criteria and are suitable for use at higher education institutions across nations.

The compilation of this work in the present book shows where we stand today and the progress that has been made in this field of research with newly developed theoretically conceptualized approaches to modeling and measurement instruments for empirical studies. It also illustrates which issues have not yet been thoroughly addressed by the – indeed very active – research community measuring student learning in higher education. Therefore, this book offers a sound basis for further research, highlighting the current challenges and future perspectives in measuring learning and learning outcomes in higher education we need to deal with in the next decades.

The contributions in this book are organized according to content and topic and are divided into three parts (Conceptual Development and Advances, Domain-Specific Student Learning Outcomes – National and International Perspectives, and Generic Student Learning Outcomes – Cross-National Comparative Approaches), giving the reader a structured overview of the wide range of student learning outcomes assessment in higher education.

The book begins with an outline of the contributions of the first thematic block, **Conceptual Development and Advances**.

In the first chapter in this part, *Research and Governance Architectures to Develop the Field of Learning Outcomes Assessment*, Hamish Coates gives an overview of research in the field and presents a conceptual approach, which indicates a possible path of development for this research area. Furthermore, he develops a framework for research and paints a picture of what assessment will look like in a decade's time.

In the second chapter, *Documenting and Improving Collegiate Learning in the United States*, Timothy Reese Cain and James C. Hearn describe the historical development and current features of learning outcomes assessment in higher education in the United States. They discuss the changing context of and increased interest in learning outcomes in the late twentieth and early twenty-first centuries, including the roles of various external actors and stakeholders. They conclude with considerations of on-campus actors at the hub of assessment processes and the evidence of changing practices that has been revealed through recent national surveys by research and volunteer groups.

In the third and last chapter in this part, *Information Management Versus Knowledge Building: Implications for Learning and Assessment in Higher Education*, Patricia A. Alexander builds a highly significant bridge between the concepts of *assessment* and *learning* in the sense of the Assessment Triangle. In the process, the author introduces another theoretical concept, which is highly relevant for learning in higher education in the information age and distinguishes between information management and knowledge building in the sense of key learning outcomes.

The second part of this book, **Domain-Specific Student Learning Outcomes – National and International Perspectives**, explores current empirical work in the

field of domain-specific learning outcomes. Research in the areas of teacher education in medicine and, as of late, in business and economics can be regarded as particularly advanced. This part gives an exemplary presentation of the work in this field and complements it with an outlook on a new European initiative for assessing domain-specific student learning outcomes.

Entrance diagnostics is an integral part of student learning outcomes assessment in higher education. The domain of medicine has had a pioneering role in this field of assessment. In their chapter *Challenges for Evaluation in Higher Education: Entrance Examinations and Beyond – The Sample Case of Medical Education*, Christiane Spiel and Barbara Schober discuss entrance diagnostics in higher education using the example of medical education. They highlight the potential limitations of typical entrance examinations for medical education; summarize the changes, concepts, and goals of entrance examinations in medical education have undergone in recent years; propose a comprehensive evaluation model for competence-based teaching; and explore implications for education.

Diagnostic competences are of vital importance not only in medicine but in teacher education practice as well. In their chapter *Teachers' Judgments and Decision Making: Studies Concerning the Transition from Primary to Secondary Education and Their Implications for Teacher Education*, Sabine Krolak-Schwerdt, Ineke M. Pit-ten Cate, and Thomas Hörstermann focus on accuracy in assessing academic achievement and potential as a core facet of teachers' diagnostic competence. Their findings regarding teachers' information processing emphasize the need to include situational and process-oriented components into models of diagnostic competence. The authors conclude with a discussion of important implications for teacher education and assessment practice.

Assessments in business and economics education are the focus of three further chapters in this part. In their chapter, *Threshold Concepts for Modeling and Assessing Higher Education Students' Understanding and Learning in Economics*, Sebastian Brückner and Olga Zlatkin-Troitschanskaia focus on student learning in economics and introduce a novel approach to modeling and measuring competences, which has great potential and is highly relevant particularly for process diagnostics over the course of studies.

While Brückner and Zlatkin-Troitschanskaia concentrate on the development of knowledge and domain-specific competences as student learning outcomes, in the next contribution, *Rescue an Enterprise from Failure: A Revolutionary Assessment Tool for Simulated Performance*, Fritz Oser, Susan Mueller, Tanja Obex, Thierry Volery, and Richard J. Shavelson present a novel approach to assessing performance in the area of entrepreneurship. They discuss measures for capturing the two competence constructs *sense of failure* and *rescue an enterprise from failure*, which are based on the initial validation results. They employ this innovative performance-oriented test instrument to measure entrepreneurial competence as a way to prevent entrepreneurial failure.

For decades, higher education in South Korea has had an excellent reputation. In the next chapter in this part, *Assessment of Economic Education in Korea's Higher Education*, Jinsoo Hahn, Kyungho Jang, and Jongsung Kim provide a comprehensive

overview of economics education and the established assessment practices in Korea, which originate in part from the United States or other countries and were adapted for the Korean context. The authors demonstrate a path toward increased international research in this area.

The part concludes with the chapter *What Do We Know – What Should We Know? Measuring and Comparing Achievements of Learning in European Higher Education: Initiating the New CALOHEE Approach* by Robert Wagenaar, who presents the new research initiative for assessing student learning outcomes in five domains at the European level. CALOHEE is developing the instrument's conditionalities for establishing cross-national diagnostic assessments, which can be applied Europe-wide. CALOHEE delivers three types of outcomes, which are outlined in this chapter: state-of-the-art reference points (benchmarks) for five academic sectors/subject areas, detailed assessment frameworks for these disciplines, and a multidimensional assessment model that does justice to the mission and profile of individual higher education institutions and degree programs.

The third part, **Generic Student Learning Outcomes – Cross-National Comparative Approaches**, encompasses contributions from the research area of interdisciplinary, generic student learning outcomes. One central innovative development in assessing generic student learning outcomes has been performance-oriented assessment. In the first chapter of this part, *International Performance Assessment of Learning in Higher Education (iPAL) – Research and Development*, Richard J. Shavelson, Olga Zlatkin-Troitschanskaia, and Julián P. Mariño present the conception of, rationale, and theoretical framework for this approach, which forms the basis of the new international research program iPAL.

The currently most widespread and most commonly used instrument for performance-based assessment of generic student learning outcomes in higher education is CLA+ by the Council for Aid to Education. It is the subsequent version of CLA, which was used in the AHELO study by the OECD. The next two chapters focus on the results of the validation studies and pilot implementations of CLA+ in many countries. In their chapter, *International Comparison of a Performance-Based Assessment in Higher Education*, Doris Zahner and Alberto Ciolfi present results from the validation of the tool for use in the United States and Italy. They conclude that this performance-based assessment enables comparative studies in higher education, and, therefore, international assessment of generic learning outcomes is feasible.

In the following chapter, *Adapting and Validating the Collegiate Learning Assessment to Measure Generic Academic Skills of Students in Germany – Implications for International Assessment Studies in Higher Education*, Olga Zlatkin-Troitschanskaia, Miriam Toepper, Dimitri Molerov, Ramona Buske, Sebastian Brückner, Hans Anand Pant, Sascha Hofmann, and Silvia Hansen-Schirra describe the adaptation and validation of CLA+ for use in Germany. The authors critically explore both the potentials and challenges of the implementation of this assessment instrument in higher education practice and the continuation of comparative studies based on the results of the validation study.

In their chapter *Validating the Use of Translated and Adapted HEIghten® Quantitative Literacy Test in Russia*, Ou Lydia Liu, Lin Gu, Prashant Loyalka, Amy Shaw, and Jane Wang present a validation study of another new assessment tool for generic learning outcomes. This test was developed in the United States and, like CLA+, has since been adapted for use in many countries. The authors provide an exemplary report on the validation of the Russian HEIghten Quantitative Literacy (QL) assessment with a representative group of students in Russia.

In addition to objective measurement of student learning outcomes, subjective measurement, for example, through self-reports, continues to be a research pillar in assessing generic student learning outcomes in higher education. In the final chapter in this part, *Comparative Study of Student Learning and Experiences of Japanese and South Korean Students*, Reiko Yamada introduces this type of assessment in the context of a comparative study of learning and experiences of students in Japan and South Korea. The findings indicate that student and faculty engagement variables appear to play important roles in the acquisition of knowledge and skills such as globalized skills, interpersonal skills, and cognitive ability.

Overall, this anthology introduces and explores important types of assessment in higher education as well as national and international developments. As a whole, this volume offers a broad overview of a relatively new field of research, which is of great significance for higher education, and demonstrates that more in-depth and extensive work in this field is necessary for developing appropriate approaches to assessing student learning in the twenty-first century. Along with other recent publications and with leading international studies in this field of research such as the AHELO study cited in these chapters, this volume offers a valuable foundation for further development of this emerging field.

This volume, which contains documentation on the current state of international research, would not have been possible without the tremendous collaboration of several researchers and experts from various disciplines and fields. We warmly thank all the authors for their active support and excellent contributions. We also thank all the reviewers and series editors for their extensive and helpful feedback and advice. Finally, we thank our graduate students at the University of Mainz for providing continuous support in preparing this volume, namely, Jennifer Fischer, Katja Kirmizakis, Mirco Kunz, and Mareike Magel.

Mainz, Germany

Olga Zlatkin-Troitschanskaia

Miriam Toepper

Christiane Kuhn

Berlin, Germany

Hans Anand Pant

Corinna Lautenbach

September, 2017

Contents

Part I Conceptual Development and Advances

- 1 **Research and Governance Architectures to Develop the Field of Learning Outcomes Assessment** 3
Hamish Coates
- 2 **Documenting and Improving Collegiate Learning in the USA** 19
Timothy Reese Cain and James C. Hearn
- 3 **Information Management Versus Knowledge Building: Implications for Learning and Assessment in Higher Education**..... 43
Patricia A. Alexander

Part II Domain-Specific Student Learning Outcomes – National and International Perspectives

- 4 **Challenges for Evaluation in Higher Education: Entrance Examinations and Beyond: The Sample Case of Medical Education** 59
Christiane Spiel and Barbara Schober
- 5 **Teachers' Judgments and Decision-Making: Studies Concerning the Transition from Primary to Secondary Education and Their Implications for Teacher Education** 73
Sabine Krolak-Schwerdt, Ineke M. Pit-ten Cate, and Thomas Hörstermann
- 6 **Threshold Concepts for Modeling and Assessing Higher Education Students' Understanding and Learning in Economics**..... 103
Sebastian Brückner and Olga Zlatkin-Troitschanskaia

7	Rescue an Enterprise from Failure: An Innovative Assessment Tool for Simulated Performance	123
	Fritz Oser, Susan Mueller, Tanja Obex, Thierry Volery, and Richard J. Shavelson	
8	Assessment of Economic Education in Korea’s Higher Education	145
	Jinsoo Hahn, Kyungho Jang, and Jongsung Kim	
9	What Do We Know – What Should We Know? Measuring and Comparing Achievements of Learning in European Higher Education: Initiating the New CALOHEE Approach	169
	Robert Wagenaar	
 Part III Generic Student Learning Outcomes – Cross-National Comparative Approaches		
10	International Performance Assessment of Learning in Higher Education (iPAL): Research and Development	193
	Richard J. Shavelson, Olga Zlatkin-Troitschanskaia, and Julián P. Mariño	
11	International Comparison of a Performance-Based Assessment in Higher Education.....	215
	Doris Zahner and Alberto Ciolfi	
12	Adapting and Validating the Collegiate Learning Assessment to Measure Generic Academic Skills of Students in Germany: Implications for International Assessment Studies in Higher Education.....	245
	Olga Zlatkin-Troitschanskaia, Miriam Toepper, Dimitri Molerov, Ramona Buske, Sebastian Brückner, Hans Anand Pant, Sascha Hofmann, and Silvia Hansen-Schirra	
13	Validating the Use of Translated and Adapted HEIghten® Quantitative Literacy Test in Russia.....	267
	Lin Gu, Ou Lydia Liu, Jun Xu, Elena Kardonova, Igor Chirikov, Guirong Li, Shangfeng Hu, Ningning Yu, Liping Ma, Fei Guo, Qi Su, Jinghuan Shi, Henry Shi, and Prashant Loyalka	
14	Comparative Study of Student Learning and Experiences of Japanese and South Korean Students.....	285
	Reiko Yamada	

Part I
Conceptual Development and Advances

Chapter 1

Research and Governance Architectures to Develop the Field of Learning Outcomes Assessment



Hamish Coates

Abstract This chapter articulates new research and governance architectures forming internationally to frame the future of learning outcomes assessment. It begins with a historical tour of the field, taking stock of the last 30 years by examining signature initiatives and geopolitical developments. The next section uses these foundations to extrapolate future technical, practical and substantive dimensions of a research framework. To give life to this framework, a picture is painted of what assessment will look like in a decade's time. The chapter's final section clarifies government arrangements, which would have the capacity to spur the kind of progress required to propel the field.

1.1 Introduction

This chapter articulates new research and governance architectures that are forming internationally and framing the future of student learning outcomes assessment. It takes stock of this establishing field and advances arrangements for spurring development. Important work must be done to reform assessment, and it is critical that researchers do not get lost into their own conversations or sidelined from the to and fro of major developments in policy and practice.

It seems common to hear in newspapers, reports and conferences that 'higher education is changing rapidly'. Higher education is becoming more central to socio-economic prosperity spurring intensification and proliferation of change (Coates 2017). Online technologies have undoubtedly changed access to much curriculum and the mechanics of much teaching, and more students than ever before move internationally to advance their academic and professional prospects. Interesting institutional variants are emerging, giving rise to new forms of governance and

H. Coates (✉)
Tsinghua University, Beijing, China
e-mail: hamishcoates@tsinghua.edu.cn

commercial opportunity. Collaborative technologies have fundamentally reformed core facets of much research and changed the way research is created, constructed and disseminated. At an aggregate level, change does indeed abound. On the ground, change is more complex and slower than espoused.

Yet as observation at almost any higher education reveals, much of how teachers and institutions assess student learning has not changed for over a century. Within changing institutional settings, new generations of faculty are of course interacting with new computing technologies to provide diverse students with information and experiences intended to prepare students with capabilities for tomorrow's world of professional work. Given such change, it is surprising that much assessment in higher education has not changed materially for a very long time and that economically and technically unsustainable practice is rife. As other chapters in this book affirm, there are an enormous number of innovative and high-quality developments, including those associated with technology advances (e.g. Shavelson et al., Chap. 10 in this volume). Still, every day around the world, students write exams using pen and paper, sitting without talking in large halls at small desks in rows. It is possible that this reflects the pinnacle of assessment, but given the lack of reflective advance over an extended period, this seems unlikely. Rather, given the enormous changes reshaping core facets of higher education, and pressures and prospects surrounding assessment, it is more likely that the 'transformational moment' is yet to come.

The assessment of higher education student learning outcomes is very important. Assessment provides essential assurance to a wide variety of stakeholders that people have attained various knowledge and skills and that they are ready for employment or further study. More broadly, assessment signposts, often in a highly distilled way, the character of an institution and its educational programmes. Much assessment is expensive, making it an important focus for analysis. Assessment shapes education and how people learn in powerful direct and indirect ways, influencing teaching and curriculum. Assessment is highly relevant to individuals, often playing a major role in defining life chances and directions (see also Cain and Hearn, Chap. 2 in this volume).

This chapter is posed at a formative time of the development of this field. As the field of higher education assessment research and reform takes shape, it is timely to step back and examine the broader developments in play, how to structure an understanding of emerging trends and what kind of governance arrangements would help transfer research into practice. This chapter tackles each of these areas, along the way articulating possible futures for assessment in higher education.

1.2 Signature Developments in Recent Decades¹

Assessment has forever played an integral role in higher education, but the most relevant antecedents for analysing contemporary development can be traced back over the past few decades (see also Cain and Hearn, Chap. 2 in this volume). This

¹The following text builds on Coates, H. (2016). Assessing student learning outcomes internationally: Insights and frontiers. *Assessment and Evaluation in Higher Education*, 41(5), 662–676 (Taylor and Francis).

section examines signature initiatives and how these have been shaped by various geopolitical developments. Clearly, taking critical stock of a field as large and diverse as higher education assessment is a useful, though challenging task—there are an enormous number of actors and initiatives, each at varying stages of maturity and diffusion. Rather than conduct an exhaustive review, it is feasible to conduct a review of a series of signature case studies, which have sought to shift policy and practice.

One broad line of development has involved specifying qualification-level outcomes. Examples include the European Qualifications Framework (European Commission (EC) 2015), the UK's Qualifications and Credit Framework (Ofqual 2015), the Australian Qualifications Framework (AQFC 2015) and the US Degree Qualifications Profile (Lumina Foundation 2015). As such titles convey, this work is developed and owned by systems, and such initiatives have served as important policy instruments for shifting beyond an anarchic plethora of qualifications, generating conversations about finding more coherence and indeed articulating the general outcomes graduates should expect from a qualification (Chakroun 2010). These system-wide structures can suffer from unhelpful collisions with fruitfully divergent local practice, but their inherent constraint is that they go no further than articulating only very general graduate outcomes (Allais et al. 2009; Wheelahan 2009). They offer little beyond broad guidelines for improving the assessment of student learning.

Going one step deeper, another line of work has sought to specify learning outcomes at the discipline level. The tuning process (González and Wagenaar 2008) is a prominent example which has been initiated in many education systems and across many diverse disciplines. Broadly, tuning involves supporting collaboration among academics with the aim of generating convergence and common understanding of generic and discipline-specific learning outcomes (see also Wagenaar, Chap. 9 in this volume). Canada adapted this work in an innovative way, focusing the collaborations around sector-oriented discipline clusters rather than education fields (Lennon et al. 2014), while in Australia a more policy-based and compliance-focused approach was deployed (Australian Learning and Teaching Council (ALTC) 2010). Such collaboration travels several steps further than qualification frameworks by engaging and building academic capacity within disciplinary contexts. Like the qualification frameworks, however, the work usually stops short of advancing assessment resources or sharing data and tends to focus instead on advancing case studies or best practice guidelines.

A slightly deeper line of development involves shared rubrics to compare assessment tasks or student performance. Moderation in assessment can play out in many ways (Coates 2010) as indeed has been the case in recent higher education initiatives. The moderation of resources has involved rudimentary forms of peer review through to slightly more extensive forms of exchange. Mechanisms have also been developed to help moderate student performance. In the United States, for instance, the Association of American Colleges and Universities (AAC&U) (Rhodes and Finley 2013) has developed VALUE (Valid Assessment of Learning in Undergraduate Education) rubrics for helping faculty assess various general skills. This has been

progressed in most recent cross institutional moderation work (AAC&U/State Higher Education Executive Officers (SHEEO) 2015; see also Cain and Hearn, Chap. 2 in this volume). The UK's external examiner system (Quality Assurance Agency (QAA) 2014) is a further example. Several such schemes have been launched in Australia, including a Quality Verification System and a Learning and Teaching Academic Standards Project, both of which involve peer review and moderation across disciplines (Marshall et al. 2013). This work travels more widely than qualification- or discipline-level specifications, for it involves the collation and sharing of evidence on student performance, often in ways that engage faculty in useful assurance and development activities. Such moderation work is limited, however, in being applied in isolation from other assessment activities and materials.

Collaborative assessments build from the developments discussed so far to advance more coherent and expansive approaches to shared assessment. As with other developments addressed here, such work plays out in myriad ways. For instance, medical progress testing in the Netherlands (Schuwirth and van der Vleuten 2012) involves the formation of shared assessment materials and administration of these in a longitudinal sense (for medical education, see also Spiel and Schober, Chap. 4 in this volume). Other assessment collaborations have focused on the development of shared tasks and analytical or reporting activities, for instance, the Australian Medical Assessment Collaboration (AMAC) (Edwards et al. 2012) and the German initiative showcased in other chapters with the umbrella title Modeling and Measuring Competencies in Higher Education (KoKoHs) (Zlatkin-Troitschanskaia et al. 2014, 2017). In 2015, the Higher Education Funding Council for England (HEFCE) funded a suite of mostly collaborative projects to assess learning gains in higher education (HEFCE 2015), and the European Commission funded a large-scale collaboration titled Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe (CALOHEE) (EC 2015). Such work is impressive as it tends to involve the most extensive forms of outcome specification, task production, assessment administration, analysis and reporting and at the same time develop faculty capacity. Work plays out in different ways, however, shaped by pertinent collegial, professional and academic factors. This can mean, for instance, that extensive work is done that leads to little if any benchmarking or transparent disclosure.

Standardised assessment is easily the most extensive form of development and would appear to be growing in scope and scale. Licencing examinations are the most long-standing and pervasive forms of assessment, though their use is cultural and they tend to be far more common in the United States than Europe (see also Cain and Hearn, Chap. 2 in this volume). Other related kinds of national effort are evident in certain countries, for instance, in Brazil (Melguizo 2015), Colombia (Shavelson et al. 2016) and the United States (Shavelson 2007; Educational Testing Service (ETS) 2014). A series of international graduate outcome tests have also been trailed in recent years, such as the OECD's Assessment of Higher Education Learning Outcomes (AHELO) (Coates and Richardson 2012), the International Association for the Evaluation of Education Achievement Teacher Education and Development Study (IEA TEDS) assessment of teachers (Braeken and Blömeke

2016), the HEIghten assessment (Liu et al. 2016; see also Liu et al., Chap. 13 in this volume), a cross national assessment of engineering competence (Loyalka 2015), and the Collegiate Learning Assessment (CLA) (CAE 2016; see also Zahner and Ciolfi, Chap. 11 in this volume). Standardised assessments are also promulgated via commercial textbooks (Pearson 2014). As implied by the term ‘standardised’ and by the external sponsorship of such work, such assessment often proceeds without engaging with academics. Though such exogenous intervention may in the longer run inject the shock required for assessment reform, it also tends to balkanise internal from external interests and has not yet been shown to have large impact on learning or teaching practice.

A variety of these practices are used by higher education providers around the world, but it must be said, in varying and inconsistent ways. Traditional assessment practices are rife in older more established universities. Such practices are baked into academic policy and procedures and more particularly into well-tenured workforces. More recently established institutions have the opportunity to leapfrog and set up more modern approaches to education design, which better express contemporary assessment ideas and practices. It is surprising that tertiary institutions are not playing a greater leading role in assessment reform given their involvement in such work over centuries. A review of the initiatives discussed above shows that most institutions are participating at the margins or in spasmodic ways, with many yet to embrace comprehensive assessment reform.

1.3 A Framework to Structure Future Trends

What do these developments over the past few decades, but particularly in recent years, tell us about the shape of things to come? This section advances a normative framework that can be used to extrapolate future technical, practical and substantive research trends. This three-dimensional framework is proposed as a mechanism for advancing principles for reforming the field of learning outcomes assessment. These dimensions are described, and the framework’s value is teased out via a number of illustrative change areas. One dimension of this framework divides change into those aspects which are substantive in nature, another which are technical in nature and the third which are practical in nature.

Substantive—policy, disciplinary and conceptual—considerations are the most significant forces shaping learning assessment. Assessment is of little use unless it is relevant to students, to policymakers, to institutional leaders and managers, to academics or to the general public. Establishing such relevance is tricky, as it involves not just identifying but also then defining what counts, and of course stakeholder interests play a role in this. Power plays a key role that manifests through the formal or informal authority of individuals or institutions. It is not uncommon to hear of conflicts regarding what should be assessed between educators, professional associations and industry or accreditation agencies. More broadly, the oligopolistic character of most established higher education systems has limited the extent to

which change has been driven by research and technological development, though appetite for research-driven change appears to be increasing with the increasingly competitive nature of higher education markets.

From a normative perspective, though evidently not always in practice, it is imperative that assessment is cogent technically. This means that assessment resources and approaches should aim to be valid and measure and report what is intended. Assessment should be reliable, which means that assessment should provide consistent measurement of the target focus area. There are a host of methods for assessing and reporting these kinds of technical properties, which of course are the focus of active scientific debates within specific communities. At a minimum, it might be expected that explicit consideration has been given to measurement considerations, but ideally a set of statistics should be provided as with professionally validated assessment instruments. Students and other key stakeholders have a right to know that assessment is producing information which pertains to people's competence in the measured area as opposed to measurement noise.

Substantive relevance and technical integrity are not sufficient to spur change in assessment. Practice is critical in that it must be feasible to collect, analyse and report data. Though institutional budgets are getting tighter, many entrenched assessment methods have high fixed costs and limited economies of scale. It is vital that more viable options are explored. Really important changes in assessment might be costly or slow to deliver. They may waste students' time and hinder learning experiences and outcomes. Indeed, such practical constraints are often claimed as impediments to progress. What matters is not just only fixed start-up costs but also ongoing costs of deployment over a prescribed time period. In building financial equations, decisions must be made about which costs are direct and indirect. A key reason for resisting change may well be that much of the cost of current assessment approaches is hidden within undifferentiated faculty roles. But the opaque nature of such costing does not make it cheap. Rather, the lack of scientific management of assessment implies all sorts of inefficiencies and scope for improvement. Of course, cost is not the only practical facet of assessment though it offers a means for summarising important decisions and uncertainties.

Each of these three dimensions plays out at varying levels. The Organisation for Economic Cooperation and Development (OECD) (2015, p. 15) notes the importance of 'distinguish[ing] between the actors in education systems: individual learners and teachers, instructional settings and learning environments, educational service providers, and the education system as a whole'. The level at which information is reported is not the same as the level at which information is collected (typically the student with assessment). Data is often collected at a lower level and then aggregated and often also combined with other data for reporting. Similarly, the interpretation level might be different again and will likely vary with the interests and concerns of stakeholders. Many current institution rankings, for instance, aggregate information on individual researcher performance and report this at the institution level, and then the information is interpreted in all sorts of ways, including in relation to fields of education. Assessment change is required for those involved in education such as students and teachers, and that change is required by broader

communities, including the general public, business and industry and people associated with planning education policy and strategy.

A series of framing ideas can be evoked from this three-dimensional framework. Substantively, it is important for assessment to be relevant or authentic to students and teachers. This often means that a diversity of assessment practice is required. At the same time, stakeholders more removed from everyday practice seek evidence which is more general in nature. Hence, a substantive idea which might be derived is that future reform should ensure that assessment is locally relevant and externally generalisable. A technical idea is that reform should advance transparency regarding the validity and reliability of assessment. The most well-designed and validated assessments are meaningless unless they are feasible to implement. Hence, a further idea for reform is that assessment must make efficient use of money and time. In terms of practice, emphasis might be placed on delivering feasible and efficient assessment to large student cohorts given tight resource constraints, whereas those more removed from the process may give more regard to the technical veracity of the evidence produced. Stereotypical remarks made by employer groups can suggest a lack of confidence in the everyday assessment by institutions of students' knowledge and skills.

Such ideas could be nuanced differently or elaborated more exhaustively, but the above formulations are sufficient to tease out the main points at play. None of the three dimensions or the kinds of ideas that they motivate are particularly surprising or controversial, though they can provoke substantial complexity and be difficult to implement. Part of the trouble arises from the conundrums provoked by attempts to harmonise or jointly optimise the dimensions in unison. Further trouble flows from negotiating the dialectic between internal and external interests. Broader considerations flow from complexities associated with generalising the assessment of complex higher-order skills across national and cultural contexts. Resolving such issues offers a chance to unlock substantial progress in the assessment of student learning outcomes. Hence, the dimensions provide a useful normative rubric against which to evaluate current progress and change dynamics and to forecast insights and frontiers for reform.

1.4 Guiding Transformation into Practice

What do the normative framework and the earlier insights regarding signature initiatives convey in terms of progress and strategies for future development? This section deploys the framework to take brief stock of the emerging field before turning to focus on potential steps ahead. A picture is painted of prospects for future development, building on earlier analyses of Coates (2014) and researchers from many different systems.

What evidence is there that the current initiatives are helping to ensure that assessment is locally relevant and externally generalisable? Large-scale qualification frameworks do very little to achieve this, but the more practice-focused initia-

tives do appear to be driving progress in this direction. The suite of programmes that invite academics to focus on organising their practice around more generalisable principles—such as the tuning process and the VALUE rubrics—provides signposts for change (see also Cain and Hearn, Chap. 2 in this volume). Simultaneously, the externally driven initiatives are themselves benefitting from technological advances in assessment (Bennett 2015) which give new insights into what assessment can look like and deliver. There would appear to be some way to go, however, in transcending the internal/external dialectic that appears to simultaneously spur and hinder progress. As well, there is much work to be done to bridge the reducing but still large gap between large-scale policy and technical development and everyday practice.

On the technical front, is progress regarding outcomes assessment advancing transparency regarding the validity and reliability of assessment? Diversified and large-scale initiatives such as KoKoHs reveal the extent of work required to validate higher education assessment (e.g. Zlatkin-Troitschanskaia et al., Chap. 12 in this volume). Larger-scale assessments such as AHELO illustrate the extent of technical validation and transparency that can be achieved. But with the workforce capacity bottlenecking progress even with funded large-scale initiatives, it is unlikely that teaching institutions will be positioned anytime soon to validate student assessment in ways that meet address psychometric standards and criteria. There is a high risk that students are assessed using insufficiently validated tasks which yield spurious information about performance and potential (Coates 2015). Inadequate information of this kind carries risk for graduates and also for institutions and countries. The need for development in this front is important, and a suggestion is given below.

Are contemporary advances helping to enhance the efficiency of assessment? Given that most large-scale initiatives tend to be relatively expensive (e.g. crude estimates of around \$10,000 per finished professionally produced multiple-choice item are not uncommon), such work is itself unlikely to be offering any intrinsic signals for how to make assessment more efficient. Through innovation, however, large-scale initiatives do carry potential to initiate new technologies and approaches and to model how new efficiencies may be achieved. The risk, of course, is that change is shaped more by factors, which are exogenous to assessment. That is, given ambiguous budget constraints and unclear technical and substantive expectations, explicitly identifiable assessment costs become a real target for savings, particularly compared with more visible staffing resources and facilities. The consequence of such disinvestment is obvious—cheaper and lower quality forms of assessment will be used that are less authentic and robust. Understanding the trade-offs linked with differential levels of direct and indirect resourcing is important, which hinges on the kind of productivity evaluations exemplified via the National Center for Academic Transformation (NCAT) course redesign initiatives (Twigg 2003). Broadly, it could be expected that in any decomposition of assessment costs it is development and implementation as opposed to the planning, analysis and reporting phases, in which new techniques carry potential to spur new economies of scale.

What does this albeit brief stocktake imply about the most fruitful areas to target reform? Where should further energy be directed to optimise substantive, technical

and practical matters and do as much as possible to address tensions associated with the internal/external dialectic? How can such energy most effectively navigate the change dynamics noted above? From the above analyses, recommendations can be made for focusing future development.

Seamless assessment tasks and processes must be prepared which can jointly serve the needs of internal and external stakeholders. This might involve production of resources, which can be shared across boundaries perhaps via adaptation to different disciplinary, professional or cultural contexts, or it might involve embedding more generic materials within local assessments. Several of the initiatives reviewed above have progressed such options. They have identified ways for harmonising the production and delivery of materials drawn from different sources, for integrating processes and for using more professionally developed materials to seed change in local practice.

Further work should be invested into techniques that engineer validity into assessment development. Rather than defer to post hoc evaluation by assessment experts, the quality of assessment is most likely to be improved by intervening earlier in the development cycle to ensure that materials exceed minimally sufficient technical standards in the first place. A specific example includes larger use of principled assessment design frameworks that help scale up assessment, so that assessment creation can be better aligned with standards, connected with learning sciences, more efficiently implemented for scaling up technologically and with conceptual frameworks suited to tailoring to local needs within a broader framework (Mislevy et al. 2011; Luecht 2013). Any such development hinges obviously on a set of accepted standards and on an effective means for bringing such standards into play. Internationally, standards do exist (e.g. American Educational Research Association (AERA), American Psychological Association (APA) and National Council for Measurement in Education (NCME) 2014; International Testing Commission (ITC) 2015), and higher education institutions have a range of means for governing the incorporation of these into academic practice. While mention of the word ‘standards’ in higher education can provoke debates about standardisation and regression to the mean (Coates 2010), there would appear to be value in progressing such work if it places a transparent floor around the quality of assessment.

Technology-assisted collaboration and delivery has an important role to play in improving practice. As Bennett (2015) conveys, by affording rethinking of task design and delivery, it also provides a frame for advancing the substantive and technical frontiers. Technology-assisted collaboration is important as this is a major means for making assessment more transparent and enhancing quality and productivity. Peer and stakeholder review of tasks helps to iron out glitches and improve authenticity. Professional capacity is developed through feedback. Technical standards could be embedded in design architectures. Sharing design and development also reduces redundancy, duplication and expensive fixed costs associated with resource production (Coates 2015). As well, with appropriate security solutions now available (Richardson and Coates 2014), it is feasible to shift from paper to online administration and reap derivative efficiencies in time and money. Of course, many platforms exist and are available from collegial or commercial sources

(e.g. Dillon et al. 2002; National Board of Medical Education (NBME) 2008; Cisco Networking Academy 2012). The key is to marry these with enterprise-level learning systems, which have scaled to ubiquity over the last decade. Put simply, such technologies should distil insights from measurement science into systems that make good assessment easy for time-poor and non-expert academics.

There is much to be considered regarding the propagation of powerful technologies into higher education. As cautioned in 2005 in relation to the rapid expansion of enterprise-level learning management systems (Coates et al. 2005), such change should not be led by technozealots who see information systems as a panacea but rather by educational leaders who can shrewdly leverage technology for change. Among other matters, it is imperative to consider the influence of systems on teaching and learning, the uncertain effects on students' engagement, the new dynamics in academic work and the organisation of teaching and the possible corporatisation of academic knowledge. As Bennett (2015) contends, most value is to be had by exploiting the sophistication of 'third-generation' technology-based assessment. This involves not just transferring traditional paper-based batch processes to computer (first generation) or incremental improvement in quality and efficiency (second generation) but fundamental redesign which ensures that assessment serves both individual and institutional needs, is informed by cognitive principles, is a medium for enabling natural interaction with rich assessment tasks and is technically enabled to support and enhance curriculum and teaching.

1.5 Governance to Spur Progress²

The preceding sections set out a normative framework and used this to frame the past and chart future assessment initiatives. What are the most effective ways of translating this work into practice? This chapter's final section articulates governance options for spurring the kind of progress required to propel the field. In particular, it advocates for more open forms of assessment, even in the most confidential and secure fields of assessment.

Analysing how best to translate research into practice can be done in a variety of ways, and the most significant considerations go to governance. New ways of designing and managing academic work, including assessment, will almost certainly require new forms of governing academic activity, power and performance (Shattock 2012). While assessment is experienced mostly as a practical educational matter, it touches many facets of higher education leadership and management. Indeed, assessment goes right to the heart of important aspects of governance such as ownership, authority and power. The risks of poorly designed or conducted governance, and the need to get governance right, show up in sectoral or organisational failures. The governance of assessment is both critical and problematic.

²The following text builds on Canny, B. & Coates, H. (2014). *Governance Models for Collaborations Involving Assessment*. Sydney: Office for Learning and Teaching.

An important feature of higher education sector is the self-accrediting status of most institutions, particularly universities. Typically, some regulatory power—typically government—delegates institutions the authority over academic programmes of course including assessment. This continues tradition that stems from the origins of the university system in mediaeval Italy and England. It does, however, create a natural tension inasmuch that universities are funded by the public purse and produce graduates for the community's benefit, but the community may not have a direct involvement in assessing the relevance and standards of education. Managing public versus private authority is an important thread running through debates about assessment reform. What spectrum of governance arrangements might be considered?

The assessment of student learning is done in myriad ways within universities. Obviously, quite a lot of assessment is done by individual academics working alone within single institutions. Alternatively, assessment can be done by groups of academics within a single institution. In each of these cases, accreditation by a government or industry authority vests power in an institution's academic board which devolves power to individual academics. The situation in practice is far more complex than suggested by this straightforward chain of command, with academics drawing on all kinds of more or less indirect and informal networks. In key respects the quality and economics of this collegial fabric are hard to beat, but at the same time, its informal and elite nature falls short. Typically, there is loose institutional oversight, academics flying solo and deployment of non-validated materials using dated practical and technical approaches.

Alternatively, assessment can be enacted and governed by groups of academics across institutions, almost invariably but not necessarily within the same discipline or professional field. Academics collaborate in this way routinely in their research work—forming collaborations and networks to design, execute and publish work. It is reasonably common for academics across institutions to share teaching, perhaps to service particular knowledge needs or to diversify teacher perspectives and student experiences. It remains far less common, however, for such collaborations to spill over into assessment. A few reasons for this have been sketched above—such as security, confidentiality and privacy—and there are doubtless others that go to individual and institutional commercial and reputational factors. Operating between institutions also carries governance implications, inasmuch as the collaboration space lies strictly beyond the jurisdiction of any single institution's reach. These implications are addressed below via the proposed academic governance model.

Assessment may also be governed from outside institutions. This work may involve academics working with third-party organisations or third-party organisations working alone. This work may take place on university campuses, or it may be outsourced to collaborative academies or statutory bodies empowered to perform specific functions for community benefit like licencing and credentialing. Such external governance is reasonably common with admissions or licencing examinations but quite rare for in-course assessment even in highly regulated fields. The delegation of assessment in this way raises even more substantial governance considerations. For instance, what are the governance arrangements of these external

organisations? Why might oversight by a private testing firm or accreditation body be preferable to that provided by a university's academic board composed of potentially hundreds of experts? How are faculty engaged? How is the authenticity of materials assured? What controls are in place to minimise the duplication of effort between faculty and external agencies?

Figure 1.1 captures this spectrum of governance arrangements sketched above. These range from individuals working alone to academics collaborating among themselves and with other agencies to fully external arrangements. As with all models, this is an abstraction, but it is helpful in clarifying the main options at play.

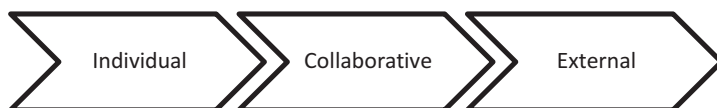


Fig. 1.1 Spectrum of assessment governance arrangements

These simplistic arrangements of course play out in an infinitely complex array of ways. Today's tertiary institutions are reforming in many areas, adopting a variety of approaches and spawning a proliferation of organisational forms of work (Coates 2017). What seems sure in all of this complexity is that the business/academic model of (in general terms) 'faculty working alone in isolation' is moving more towards 'faculty playing a role in a broader team'. Where new forms of academic infrastructure are emerging, these are typically affiliated or owned by existing institutions. In essence, renovating/improving rather than replacing existing academic arrangements seems to be the most common change underway. Shifting to more collaborative forms of assessment governance would do well to reflect such transition.

The 'sharing economy'—or perhaps in higher education the 'collaboration economy'—is reshaping many facets of economic and social life, and higher education is no exception. Rather than goods and services being created and used by individuals in isolation, teachers and learners are collaborating via advanced online systems to generate new ways of doing education. Teachers and institutions are collaborating on curriculum production, learners are collaborating on assignments and open admissions, provision and credit recognition are touching basic notions of the qualification. As signalled at the outset, the assessment component of education has been one of the most resistant areas to adapt to the changing environment. In many areas assessment is closely tied via content and implementation to local educational settings. It has obvious security, confidentiality and privacy aspects. As the tool for evaluating individual performance, it also helps measure the quality of programmes and institutions and through this carries reputational and commercial implications. For these and other reasons, assessment would appear to be one of the final frontiers in the contemporary unbundling of higher education.

By taking stock of recent signature developments and painting a picture of future practice, this chapter has advanced a framework for thinking through the growth of the field and productive new forms of governance. Technology almost certainly will

play a role in changing practice, but it is essential that effective governance architectures are in place. This chapter has advocated reform by strengthening and augmenting rather than replacing traditional collegial arrangements. Most particularly, the paper has advocated for the value of moving to more collaborative kinds of governance. Even—and perhaps especially—the most high-stakes assessment needs to become more open to improve.

References

- Allais, S., Young, M., & Raffe, D. (2009). Introduction. In S. Allais, D. Raffe, R. Strathdee, L. Wheelahan, & M. Young (Eds.), *Learning from the first qualifications frameworks*. Geneva: International Labour Office.
- American Educational Research Organisation (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational & psychological testing*. Washington, DC: American Psychological Association.
- Association of American Colleges and Universities (AAC&U), & State Higher Education Executive Offices (SHEEO). (2015). *MSC: A multi-state collaborative to advance learning outcomes assessment*. Accessed 20 Sept 2016 from: <http://www.sheeo.org/projects/msc-multi-state-collaborative-advance-learning-outcomes-assessment>
- Australian Learning, & Teaching Council (ALTC). (2010). *Learning and teaching academic standards project final report*. Sydney: Australian Learning and Teaching Council.
- Australian Qualifications Framework Council (AQFC). (2015). *Australian qualifications framework*. Accessed 1 Sept 2016 from: <http://www.aqf.edu.au>
- Bennett, R. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39(1), 370–407.
- Braeken, J., & Blömeke, S. (2016). Comparison of beliefs across countries: Dealing with measurement invariance and local dependence using Bayesian elastics. *Assessment and Evaluation in Higher Education*, 41(5), 733.
- Canny, B., & Coates, H. (2014). *Governance models for collaborations involving assessment*. Sydney: Office for Learning and Teaching.
- Chakroun, B. (2010). National qualification frameworks: From policy borrowing to policy learning. *European Journal of Education*, 45(2), 199–216.
- Cisco Networking Academy. (2012). *Advancing assessment with technology*. Accessed 2 Dec 2016 from: http://www.cisco.com/c/dam/en_us/training-events/netacad/downloads/pdf/NetAcadPOV.pdf
- Coates, H. (2010). Defining and monitoring academic standards in Australian higher education. *Higher Education Management and Policy*, 22(1), 1–17.
- Coates, H. (2014). *Higher education learning outcomes assessment: International perspectives*. Frankfurt: Peter Lang.
- Coates, H. (2015). Assessment of learning outcomes. In R. Pricopie, P. Scott, J. Salmi, & A. Curaj (Eds.), *Future of higher education in Europe. Volume I and volume II*. Dordrecht: Springer.
- Coates, H. (2016). Assessing student learning outcomes internationally: Insights and frontiers. *Assessment and Evaluation in Higher Education*, 41(5), 662–676.
- Coates, H. (2017). *The market for learning: Leading transparent higher education*. Dordrecht: Springer.
- Coates, H., & Richardson, S. (2012). An international assessment of bachelor degree graduate's learning outcomes. *Higher Education Management and Policy*, 23(3), 51–69.

- Coates, H., James, R., & Baldwin, G. (2005). A critical examination of the effects of learning management systems on university teaching and learning. *Tertiary Education and Management, 11*, 19–36.
- Council for Aid to Education (CAE). (2016). *Collegiate Learning Assessment (CLA)*. New York: CAE.
- Dillon, G. F., Clyman, S. G., Clauser, B. E., & Margolis, M. J. (2002). The introduction of computer-based case simulations into the United States medical licensing examination. *Academic Medicine, 77*(10), 94–96.
- Educational Testing Service (ETS). (2014). *Proficiency profile*. Accessed 1 Sept 2014 from: www.ets.org/proficiencyprofile/about
- Edwards, D., Wilkinson, D., Coates, H., & Canny, B. (2012). *The Australian medical assessment collaboration: Developing the foundations for a national assessment of medical student learning outcomes*. Sydney: Office of Learning and Teaching.
- European Commission (EC). (2015). *European qualifications framework*. Accessed 1 Sept 2016 from: https://ec.europa.eu/ploteus/search/site?f%5B0%5D=im_field_entity_type%3A97
- González, J., & Wagenaar, R. (2008). *Universities' contribution to the Bologna process: An introduction*. Bilbao: Universidad de Deusto.
- Higher Education Funding Council for England (HEFCE). (2015). *£4 million awarded to 12 projects to pilot measures of learning gain*. Accessed 1 Sept 2016 from: <http://www.hefce.ac.uk/news/newsarchive/2015/Name,105306.en.html>
- International Testing Commission (ITC). (2015). *Guidelines*. Accessed 10 Sept 2016 from: <https://www.intestcom.org/page/5>
- Lennon, M. C., Frank, B., Lenton, R., Madsen, K., Omri, A., & Turner, R. (2014). *Tuning: Identifying and measuring sector-based learning outcomes in postsecondary education*. Toronto: Higher Education Quality Council of Ontario.
- Liu, O. L., Mao, L., Frankel, L., & Xu, J. (2016). *Assessing critical thinking in higher education: The HEIghten™ approach and preliminary validity evidence*. Princeton: Educational Testing Service.
- Loyalka, P. (2015). *Initial results from an international study of student learning in higher education: China, Russia, and the U.S.* Keynote paper presented at the 2nd symposium of learning science and online education, China, Beijing.
- Luecht, R. M. (2013). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice*. New York: Routledge.
- Lumina Foundation. (2015). *The degree qualifications profile*. Accessed 1 Sept 2016 from: <http://www.luminafoundation.org/files/resources/dqp.pdf>
- Marshall, S., Henry, R., & Ramburuth, P. (2013). *Improving assessment in higher education: A whole of institution approach*. Sydney: New South Books.
- Melguizo, T. (2015). *Are students gaining general and subject area knowledge in university? Evidence from Brazil*, Higher Education.
- Mislevy, R., Haertel, G., Yarnall, L., & Wentland, E. (2011). Evidence centered task design in test development. In C. Secolsky (Ed.), *Measurement, assessment, and evaluation in higher education*. New York: Routledge.
- National Board of Medical Examiners (NBME). (2008). *Primum® Computer-based Case Simulations (CCS)*. Philadelphia: NBME.
- Ofqual. (2015). *Qualifications and credit framework*. Accessed 1 Sept 2016 from: <https://www.gov.uk/government/organisations/ofqual>
- Organisation for Economic Cooperation and Development (OECD). (2015). *Education at a glance*. Paris: OECD.
- Pearson. (2014). *MyEconLab*. Accessed 27 Aug 2014 from: www.pearsonmylabandmastering.com/northamerica/myeconlab
- Quality Assurance Agency (QAA). (2014). *The UK quality code for higher education*. Retrieved from: www.qaa.ac.uk

- Rhodes, T., & Finley, A. (2013). *Using the VALUE rubrics for improvement of learning and authentic assessment*. Washington: AAC&U.
- Richardson, S., & Coates, H. (2014). Essential foundations for establishing equivalence in cross-national higher education assessment. *Higher Education*, 68(6), 825–836.
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). The use of progress testing. *Perspect Medical Education*, 1(1), 24–30.
- Shattock, M. (2012). University governance. *Perspectives: Policy and practice in higher education*, 16(2), 56–61.
- Shavelson, R. J. (2007). *A brief history of student learning assessment: How we got where we are and a proposal for where to go next*. Washington, DC: Association of American Colleges and Universities.
- Shavelson, R. J., Domingue, B. W., Mariño, J. P., Molina-Mantilla, A., Morales, J. A., & Wiley, E. E. (2016). On the practices and challenges of measuring higher education value added: The case of Colombia. *Assessment and Evaluation in Higher Education*, 41(5), 695–720.
- Twigg, C. A. (2003, September/October). Improving learning and reducing costs: New models for online learning. *EDUCAUSE Review*, 38(5), 28–38.
- Wheelahan, L. (2009). From old to new – The Australian qualifications framework. In S. Allais, D. Raffe, R. Strathdee, L. Wheelahan, & M. Young (Eds.), *Learning from the first qualifications frameworks*. International Labour Office: Geneva.
- Zlatkin-Troitschanskaia, O., Kuhn, C., & Toepper, M. (2014). Modelling and assessing higher education learning outcomes in Germany. In H. Coates (Ed.), *Advancing higher education learning outcomes*. Frankfurt: Peter Lang.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Toepper, M., Lautenbach, C., & Molerov, D. (2017). Valid competency assessment in higher education: Framework, results, and further perspectives of the German Research Program KoKoHs. *AERA Open*, 3(1), 1–22.

Chapter 2

Documenting and Improving Collegiate Learning in the USA



Timothy Reese Cain and James C. Hearn

Abstract This chapter examines the expansion and current features of learning outcomes assessment in the US higher education. It begins with an overview of the diverse and stratified higher education system, including the multiple layers of state and federal influence but lack of centralized tightly linked control mechanisms. It describes the changing context of, and increased interest in, learning outcomes in the late twentieth and early twenty-first centuries, including the roles of various external actors. In so doing it highlights the regional and national accrediting bodies that serve as gateways to federal financial aid funds and therefore serve as important means of affecting change. Although assessment is important for institutional accountability, its true value lies in its as-yet-unrealized potential to fundamentally improve the teaching and learning at the campus level. As such, the chapter concludes with considerations of on-campus actors at the heart of assessment processes and the evidence of changing practices that have been revealed through recent national surveys by research and voluntary membership groups.

2.1 Introduction

Over the past four and a half decades, higher education stakeholders in the USA have become increasingly concerned about what students know and what abilities they possess as a result of their college educations. Pushed by external demands for accountability and internal concerns about learning, outcomes assessment has become an important consideration at institutional, state, and federal levels—along with equity, safety, and completion, it is one of the most important concerns about college students in the second decade of the twenty-first century. At the same time, both the idea and practice of assessment have remained controversial. Many question whether it is even possible to measure the most important student outcomes and whether attempts to do so might change priorities and affect

T. R. Cain (✉) · J. C. Hearn
University of Georgia, Athens, GA, USA
e-mail: tcain@uga.edu; jhearn@uga.edu

activities in ways that are detrimental to student learning. Some link the practice to a growing corporatization of higher education, where faculty authority is being replaced by administrative oversight. Others are more confident that we can document learning but worry that we largely do so to meet accountability requirements, rather than using the results to effect change in student learning, thereby fulfilling assessment potential. Yet, even with these concerns, student learning outcomes remain vital to understanding and improving American colleges and universities.

In this chapter, we consider student learning outcomes assessment in the USA, emphasizing the key actors and the major issues with which they are engaging. We begin with a brief overview of the diversified system of American higher education, highlighting the different governance structures, institutional types, and current concerns that provide context for understanding how assessment has been shaped and operates. We then chart a history of the assessment movement in the USA, from its beginnings through the modern era, emphasizing the evolution of what was once seen as a fad to what has become a central facet of accountability mechanisms and key, if contested, campus concern. We turn next to three major external constituent groups that are influencing assessment: governmental actors; accrediting agencies that mandate that colleges and universities report on their students' learning; and the membership associations and foundations that are driving much of the conversation about assessment. The heart of assessment is what happens on college campuses, where students, faculty, administrators, and others interact—or in many cases do not—around learning outcomes and the ways to improve them, so we rely on large national surveys to consider on-the-ground campus practices. Finally, we conclude with key issues, challenges, and remaining questions.

2.2 Overview of American Higher Education

From an international perspective, US tertiary education is distinctive in several respects. Among its most prominent defining characteristics are its size, its differentiation, its marketized control and coordination, and the holistic missions and activities of its individual institutions.

Size

US higher education is approaching what sociologist Martin Trow (2007) termed “universal access.” Over 60% of all secondary school graduates enter postsecondary education directly after graduation (National Center for Higher Education Management Systems 2016), but substantial and growing numbers of students enroll many years after secondary school graduation. In fall 2014, 4665 institutions enrolled over 20 million students, representing approximately 1 out of every 12 US adults (*Chronicle of Higher Education Almanac* 2016; U.S. Census Bureau 2016). Some national leaders, however, express consternation over the USA losing its global lead in tertiary attendance and graduation rates (<https://www.whitehouse.gov/issues/education/higher-education>), and there are signs that the public's declining faith in the quality and value of college attendance may threaten further enrollment growth (Public Agenda 2016).

Differentiation

The USA has several distinctive tertiary sectors, several forms of faculty employment, and several forms of student enrollment. Popular images of college have traditionally featured a pastoral setting, a veteran professor embodying traditional academic culture, and a youthful student enjoying a campus life split between studies, parties, and sporting events. Such images inadequately reflect contemporary realities. Fewer than a fifth of US students attend the familiar and iconic institutions featured in popular imagery: the large public (e.g., University of California, Los Angeles (UCLA), University of Michigan) and private (e.g., Stanford, Harvard) research universities and the idyllic liberal arts colleges (e.g., Williams, Amherst). Instead, nearly one-fourth attend colleges offering only the associate degree (designed to be attained in the equivalent of 2 years of full-time study and mainly offered by community colleges), and the remaining majority attend religiously affiliated institutions, vocationally specialized institutions, institutions offering baccalaureate and master's degrees but not doctorates, and colleges for special and underserved populations (*Chronicle of Higher Education Almanac* 2016).

Across these varied sectors spread further distinctions in controlling authority. The nation's colleges and universities fall into three categories: public (1644 institutions directly supported in part by government allocations), private not-for-profit (1731 institutions), and for-profit (1290 schools). Although the majority of tertiary institutions in the USA are not public, most students attend public institutions.

The structural differentiation is paralleled by employment differentiation among academic staff within institutions. In the mid-1900s, many US colleges institutionalized high levels of job security (tenure) as a path to protecting the academic freedom of scholar-teachers while also serving as a form of noneconomic compensation. Formalized tenure grew for many years, but by the 1970s, the expansion of higher education brought with it renewed reliance on nontenure-line and part-time faculty, and colleges and universities have increasingly moved away from relying on a heavily tenured workforce (Hearn and Deupree 2013; Schuster and Finkelstein 2006). Not including the graduate students who undertake significant teaching responsibilities at doctoral universities, two-thirds of all US academic staff now work on nontenure-line contracts (Finkelstein et al. 2016). And among that new majority, there is extensive variation in the nature of appointments as to full-time vs. part-time employment, governing rights, salaries, contract duration, and the like. With only a minority of all faculties unionized and with the absence of a strong national guild, faculty appointments continue to evolve in disparate ways.

Differentiation further extends into the nature of tertiary students in the USA. As with faculty, substantial numbers of students participate part-time (38.2% in fall 2014, according to the US Department of Education (2014)). In addition, substantial numbers attend only after having spent a few years in the workforce after high school graduation. That pattern of delayed enrollment contributes to growing numbers of older and "nontraditional" learners within the system (nearly a third of all tertiary students fit that pattern in fall 2014, according to the *Chronicle of Higher Education Almanac* (2016)). There is also variation in students' residential status, with many still living on campuses but growing numbers living in off-campus housing or participating in distance or online classes (*Chronicle of Higher Education Almanac* 2016).

Stratification

Institutions in the USA may be arrayed along a continuum of perceived stature and prestige based in the academic preparation of their students and the intensity of their research enterprise. In some ways, this stratification reflects governmental choices: state governments target some public institutions for greater funding based on their more intensive graduate education and research focus, their more extensively educated faculty, and their higher student selectivity, while the federal government targets its research funding on a relatively small set of institutions with the strongest research faculty and scientific facilities. These governmental choices on admission selectivity and research funding contribute to public perceptions of quality, and together these forces are reflected in popular ranking systems such as those of the *US News and World Report* (Bastedo and Bowman 2010, 2011). Governmental choices and public perceptions of quality tend to reinforce each other over time. Institutions attracting the most academically prepared students and securing the most research funding benefit from public and professional perceptions of quality, which in turn send more students and research support their way. Turning to a biblical analogy, sociologists (e.g., Trow 1984) have labeled this a “Matthew effect” at work in higher education: the rich get richer while the poor get poorer. Whatever one’s chosen metaphor, it is clear that the stratification in US tertiary education is as much socially constructed as governmentally dictated.

This institutional stratification has important equity implications. The socioeconomic characteristics of students are strongly associated with their placement within the various postsecondary sectors in the USA. That is, the system appears most open and egalitarian when one defines tertiary education broadly, to include 2-year and for-profit institutions. To the extent one limits one’s attention to the most prestigious universities and colleges, students from lower socioeconomic backgrounds are significantly underrepresented (Bastedo and Jaquette 2011; Hearn and Rosinger 2014).

Marketized Control and Coordination

No national ministry controls US tertiary education. Education is unmentioned in the US Constitution and is historically devolved to a state-level responsibility. That said, the role of the federal government in the enterprise is limited largely to funding and legal concerns, and the 50 states vary greatly in the ways they approach public higher education. Some states, such as North Carolina, Wisconsin, and Georgia, feature highly centralized governing boards that can influence academic programming and staffing, differentiate institutional missions, structure and fund developmental and remedial education policies, devise transfer and articulation policies and practices, mandate admissions standards, and design and control distance and online learning options (Hearn and Holdsworth 2002). Other states, however, employ only relatively weak coordinating or planning units. This remarkable variation in governing arrangements makes generalizations across states impossible.

What can be said confidently is that tertiary governance lodges most substantively at the levels of the chosen boards of state higher education systems and institutions, the chosen boards of private institutions, and the leaders and

faculty of individual colleges and universities. Under Burton Clark's (1979) typology of forms of coordination in higher education, the USA relies more heavily than most other nations on market and professional forces, trusting the decisions of local actors and students more fully than the authority of removed bureaucratic or political actors. That is not to say that bureaucratic and political coordination are trivial, however. Indeed, their influence is arguably growing in the new century (McLendon and Hearn 2009).

Holism

Unlike many tertiary systems elsewhere, institutions in the USA tend to espouse wide-ranging missions and engage in diverse activities. For example, research and teaching are often both included in the missions of individual universities—free-standing institutes for academic research are relatively rare. Similarly, most public colleges in the USA incorporate public service into their missions, a commitment that can be dated to the nineteenth century. Further, the student experience at traditional place-based campuses often involves voluntary social organizations such as fraternities and sororities, intramural and intercollegiate athletics, and a wide variety of other extracurricular activities. Another indication of the breadth of the student experience is the commitment of many institutions to encourage their students to live in on-campus residence halls featuring a wide variety of academic and non-academic activities. Participation in many of these activities is linked to an array of student outcomes, although not always explicitly tied to *learning* outcomes. As noted earlier, the shifting attendance patterns and student populations are upending some of these traditional elements.

2.3 History of Learning Outcomes Assessment in the USA

Concern about the assessment of student learning has a long history in American higher education, with Shavelson (2007; see also Shavelson et al., Chap. 10 in this volume) identifying four distinct eras of outcomes assessment as related to testing. The first third of the twentieth century saw the beginning and rise of objective standardized testing, in part through the efforts of the Carnegie Foundation for the Advancement of Teaching. From 1933 to 1947, comprehensive examinations of achievement took precedence, many of which were linked to the widespread concern about and reform in general education in American colleges and universities. Likewise important was the development and use of the Graduate Record Examination (GRE), which first tested content knowledge that could be acquired in a typical college and then, beginning in 1949, shifted to measuring general reasoning. American higher education experienced tremendous growth in the number of students and institutions in the decades following the war and with it came the growth in testing companies such as Educational Testing Service (ETS). These testing companies promoted and sustained the widespread use of standardized and cost-efficient exams that could efficiently capture data through multiple-choice formats (e.g., Liu et al., Chap. 13 in this volume). They were soon joined by more

holistic exams that sought to overcome the shortcomings of multiple-choice exams and provide truer measures of key outcomes of college, including critical thinking and communication abilities. These, though, proved costly and difficult to reliably score. Shavelson identified the period from 1979 through the modern era as a period of accountability, in which state and federal stakeholders demanded that colleges and universities begin to demonstrate their value and effectiveness. With the legislative prompt and internal concerns over the value of multiple-choice tests, new ways of assessing student learning that called on students to construct responses and demonstrate more complex learning became key considerations.

This modern era of outcomes assessment in the USA includes but extends beyond the testing considerations that Shavelson (2007) outlined. It had small but significant beginnings in institution-based assessment programs in the 1970s, most famously at Alverno College, a private woman's college in Milwaukee, Wisconsin, that reshaped its entire curriculum around eight core competencies, now known as abilities: communication, problem solving, social interaction, effective citizenship, analysis, valuing, developing a global perspective, and esthetic engagement. With a centralized assessment center and ongoing and integrated assessment throughout students' college careers, Alverno was and is a pioneer in the field and exemplar of what could happen when an institution committed itself to using assessment to improve student learning (Allen 2016; Sims 1992). So, too, was Northeast Missouri State University, which fully committed to student assessment in the early 1980s as part of strategic institutional change (Gaston 2014). In the 1980s, faculty across the nation were beginning to engage in conversations leading to what would come to be known as the Scholarship of Teaching and Learning (SOTL) after the publication of Ernest Boyer's (1990) *Scholarship Reconsidered*. SOTL was and is a sustained effort designed to turn investigative lenses toward improving pedagogy, which is distinct from but contributed to internal assessment efforts.

Just as important was a series of reports from both inside and outside of the academy that questioned the efficacy of undergraduate education in the USA and linked efforts for reform to assessing outcomes (e.g., National Institute of Education (NIE) 1984; Association of American Colleges & Universities (AACU) 1985). Some of these came from and spoke directly to state policy-makers—such as the National Governors Association's (1986) *Time for Results: The Governors' 1991 Report on Education*—and pushed a new wave of state-level accountability efforts in the 1980s and early 1990s. These included a turn to outcomes assessment, with over half of the states beginning some sort of assessment policy initiative in the period. By the end of the century, external accreditors—membership organizations charged with assuring institutional quality and serving as gatekeepers to federal student aid funds—replaced states as the primary external drivers of learning outcomes assessment. Accreditors' roles took on further importance in 2005–2006 when the Commission on the Future of Higher Education, commonly referred to as the Spellings Commission after then US Department of Education Secretary Margaret Spellings, sought significantly increased oversight and accountability of American tertiary education. The Commission actively debated the value and accomplishments of accreditation and whether it was or could be up to meeting the challenges of

providing needed quality assurance for colleges and universities. In its final report, the Commission noted its concern with declining quality and outcomes and argued

We believe that improved accountability is vital to ensuring the success of all the other reforms we propose. Colleges and universities must become more transparent about cost, price, and student success outcomes, and must willingly share this information. Student achievement, which is inextricably connected to institutional success, must be measured by institutions on a “value-added” basis that take into account students’ academic baseline when assessing their results. (U.S. Department of Education 2006, p. 4)

Accreditation ultimately was retained as the key oversight mechanism but with new impetus to focus on the outputs of education, including as related to learning, and with greater responsiveness to the public. The Spellings Commission prompted further action in the area of outcomes, most notably the creation of the Voluntary System of Accountability (VSA). A project undertaken by the American Association of Public and Land-grant Universities (APLU) and the American Association of State Colleges and Universities (AACSB), the VSA was designed to provide a unified format and web-portal through which member colleges and universities could share their outcomes information. Although successful in addressing the political crisis of potentially significant increases in federal oversight, in its earliest iteration, it actually did little to increase public consumption of outcomes information or improve learning outcomes on college and university campuses (Ikenberry and Kuh 2015).

In the second decade of the twenty-first century then, we have very real internal concerns about student learning and the ways that it might be improved conjoined with major external drivers of learning outcomes assessment, frequently enacted through an accountability lens. Those accountability demands are so dominant that they have come to shape both how assessment is undertaken and how it is perceived. The substantial pushes that accreditors have given to make institutions undertake assessment activities has been conflicted. Activities have certainly increased in ways that they would not have otherwise, yet they have also met substantial resistance from stakeholders, particularly faculty, who view the efforts as counterproductive, an encroachment on both institutional autonomy and academic freedom, and a fadish legacy of the total quality management movement of the 1990s. Moreover, the accountability and compliance framing has alienated many and led institutions to emphasize reporting student outcomes rather than improving them (Kuh et al. 2015).

2.4 External Actors and Drivers of Assessment

2.4.1 *Government Actors*

The federal and state governments in the USA have very little direct power or control over day-to-day academic practices and outcomes in higher education. Instead, they seek to affect students’ learning outcomes via more indirect paths, using the leverage of resource allocations, voter-based political power, and informal influences. That said, there can be little debate that, since the 1980s, and especially

in the past two decades, the federal government and the states have emphasized the necessity of improving student outcomes, arguably at the expense of earlier emphases on expanding educational access and equity (e.g., Hillman 2016). At the federal level, efforts have focused mainly on leveraging student financial aid, publicizing indicators of college success with students, and using the power of persuasion to influence various stakeholders. At the state level, initiatives to improve college outcomes have been more diverse, and probably more impactful, than federal efforts, though at both levels many of the efforts focus on outcomes other than learning, such as graduation and employment rates (see also Coates, Chap. 1 in this volume). These two levels of government efforts will be discussed in turn.

The Federal Government

In student financial aid, the federal government has gradually but consistently increased its use of the leverage provided by its massive grants and loan programs to incentivize students' "satisfactory academic progress" through their college years. Total years of eligibility for student aid have been limited, and minimal credit accumulation has been required for students to maintain eligibility. In addition, penalties and potential program exclusion have been imposed on schools with poor records on students' graduation and loan repayment. Advocates for learning outcomes argue that merely passing classes and earning degrees are poor measures of actual student learning.

Vaguer, but perhaps equally powerful, has been the federal government's use of the power of suasion, or what the early twentieth century US President Theodore Roosevelt termed the "bully pulpit." That is, presidential administrations have employed the visibility and influence of the executive branch to spur improvements in the quality of undergraduate education. A highly visible and controversial example was the previously alluded to 1984 publication of *Involvement in Learning*, a report from the federal education department that excoriated undergraduate institutions' inadequate attention to student learning outcomes (NIE 1984).

Although those earlier reports employed data in the service of their argument, more recent "bully pulpit" efforts have sought to place data at the center of their arguments. Too often, however, actual student learning has eluded consideration. In an effort to make students' college-going choices more informed, and costs and benefits more transparent, the Obama administration worked to highlight the financial, academic, and safety performance of every college participating in federal programs. Most visibly, the government instituted an online "College Scorecard" for students and families (<https://collegescorecard.ed.gov>). When announced in 2013, the project was to be a massive rating system linking college scores to financial aid eligibility. Yet, it was met with severe criticism over numerous concerns including poor quality data and unintended consequences (e.g., Field 2013). Most relevant here is that actual learning was not part of the rating scheme, prompting the higher education membership association that is perhaps most involved in that arena, AACU, to call on President Obama to avoid language about value and worth and rename it "Selected Indicators on Cost and Completion" (AACU 2013). The scheme was ultimately abandoned in favor of a more modest search tool, but one that likewise avoids all mention of learning outcomes.

In concert, the Obama administration emphasized the nation's lagging competitive position internationally in educational access and graduation outcomes, and worked consistently to use both fiscal resources and "the pulpit" to increase degree attainment and workforce preparedness, especially through the nation's 2-year community colleges. The administration's 2020 initiative, for example, called for the country to once again achieve the highest proportion of college graduates in the world by 2020. It sought to do so by ensuring affordable, high-quality college availability. Announced soon after Obama's election, the 2020 initiative spurred numerous supportive efforts by professional and institutional associations, foundations, collaboratives, and state governments, some of which were captured in AASCU's (2011) useful summary of supportive private initiatives paralleling the White House goals; most of those initiatives continue today. Yet, while access and completion are laudable and needed national goals, emphasizing them without including substantial attention to learning was problematic.

State Governments

States' accountability efforts in recent years have focused mainly on efficiency, cost control, and affordability, but states have also initiated a diverse array of initiatives directly or indirectly aimed at improving the outcomes of students, especially students in public institutions. Most often, these are undertaken outside of a learning outcomes framework, focusing either on input to higher education or on outcomes that are more easily measured than learning. Still, roughly half of US states do require some form of accounting for learning, whether through standardized exams or other, often institutionally driven, means (Kinzie et al. 2015).

The most obvious way states can influence learning is through funding systems. While historically popular funding systems using enrollment-based formulas privilege institutions with the strongest raw enrollment numbers, emerging incentive-based systems target rewards on specific kinds of student outcomes. Performance or outcomes-based funding of institutions is the prime example: a majority of states in recent years have moved to score institutions' performance on such outcomes as graduation numbers, graduation rates, and job placements, and in doing so have, at least in theory, heightened institutions' attention to those results of their efforts (Hearn 2015). Another example of incentives-based funding is state investment in merit-based scholarships, which are awarded only to students maintaining strong academic performance. The state of Georgia initiated a lottery-funded merit program in the early 1990s, and numerous other states have followed suit in succeeding years (Doyle 2010).

A second way states can influence learning is through targeted budget allocations. Capital spending for buildings, for example, selectively expands educational capacity and, in theory, educational quality. New classroom buildings can enrich learning, and the expansion of study spaces and other improvements in the "quality of life" for undergraduates may help ensure students persist to earn their degrees.

Less directly, states' funding sources affect learning outcomes. To the extent states rely heavily on current revenues to fund higher education, and keep few or no reserves for difficult economic times, state funding for higher education can be extraordinarily cyclical (Hovey 1999). That cyclicity can push institutions toward

restricted spending on faculty, increased class sizes, decreased services, and so forth. Similarly, to the extent states rely on lotteries to fund merit-based student aid (a frequent choice), programs may be cut when lottery revenues decline.

Beyond funding, states can and do influence student learning through their governing arrangements. As noted above, the 50 states vary greatly in the ways they approach public higher education, ranging from highly centralized consolidated governing boards, such as in North Carolina, to relatively weak coordinating or planning units in Michigan, Colorado, and several other states (<http://www.nchems.org/psgov/>).” In the most empowered settings, states can influence academic programming and staffing, differentiate institutional missions, structure and fund developmental and remedial education policies, devise transfer and articulation policies and practices, mandate admissions standards, and design and control distance and online learning options (Hearn and Holdsworth 2002). In some ways, these are among the powers most directly vested in state-level officials in the USA, but, again, the diversity of governing arrangements makes generalizations impossible.

Institutions’ Special Status Under Federal and State Law

Chartered colleges and universities claim special legal status, and that status can potentially benefit student learning. Because of their exemption from property, sales, and income taxes under both federal and state law, institutions are able to maintain resources enabling them to provide additional learning-related services to students and faculty and staff. Colleges and universities also benefit from this status by qualifying to borrow on the open market at lower interest rates. This can facilitate the provision of high-quality facilities for student learning.

Tensions Between Intervention and Autonomy

The increasing national concern over postsecondary quality has translated at both the federal and state levels into tensions between the nation’s historic allegiance to institutional autonomy and its growing demands for tougher public accountability (Newman 1987; Schmidlein and Berdahl 2011). In this context, both federal and state governments fluctuate from serving as *interveners*, via centralizing authority, coordinating, and regulating, to serving as *encouragers*, via goal-setting, planning, creating task forces, and establishing incentive systems (Hearn and Holdsworth 2002).

At both levels, one can make the argument that the intervener role is growing: witness, for example, recent federal efforts to cut funding to poorly performing institutions, especially schools in the for-profit sector (Smith 2015). Interestingly, however, one can also make the argument that the encourager role is growing: witness, for example, the increasing levels of discretion being allowed institutions to achieve goals set under outcomes-based funding, through the increasing freedom being provided institutions to set their own tuition levels, and the like (McLendon and Hearn 2009). The simultaneous persuasiveness of both those arguments serves nicely to foreground the challenges of neatly characterizing governmental influences on learning outcomes in the USA.

While the trend of governments’ growing emphasis on outcomes is clear, the implications for student learning are not. Trends toward lowered government

support, increased marketization, and revenue diversification in US higher education may actually contribute to de-emphasizing student learning. After all, if the route to improving student graduation rates and other academic outcomes runs through tighter admission standards and more rigorous and punishing standards for advancement, then that route heads in the opposite direction from political leaders' goal of expanding the supply of well-prepared graduates. Alternatively, if the goal is expanding the number of college graduates, the quality of academic attention each student receives is threatened by tightening public funding in the USA. In sum, in a context of lessened governmental spending, one can choose increasing the quantity of graduates or improving their individual learning, but choosing both is difficult. State governors and legislators, in particular, may be asking the impossible in demanding that public institutions produce more qualified graduates with less public funding.

However, if one characterizes recent government initiatives, it should be remembered that none of the governmental activities and arrangements discussed above directly connects to student learning per se. No federal or state government in the USA is making a significant effort to measure undergraduates' achievement or competency outcomes at the student level. Instead, assessing learning outcomes is left to other parties, prominently including accreditors, individual institutions, and professional associations.

2.4.2 Extra-governmental Organizations

Governmental actors have key roles and, at times, have emphasized learning outcomes in ways that have pushed institutional engagement, but other actors often play more significant roles both in response to larger governmental actions and in trying to direct them. Among the most significant of these are the regional and programmatic accreditors, higher education and disciplinary membership associations, and philanthropic organizations.

Regional and Program Accreditors

Regional accreditors—the associations that accredit almost all not-for-profit institutions and some for-profit institutions—have been involved in learning outcomes assessment since the mid-1980s and took on a prominent role beginning in the 1990s (Ewell 2002; Powell 2013). In the aftermath of the Spellings Commission and through reauthorization of federal legislation, the higher education accrediting bodies have maintained that important role in shaping conversations about and prompting institutions to enact learning outcomes assessment. Through their service as the gateway to federal student aid funds—without accreditation, institutions are ineligible to participate in the massive federal student aid program, a virtual necessity for institutional survival—as well as larger advocacy and educational efforts, the accreditors have helped pushed US tertiary education into more fully engaging with assessing student learning, though often with a compliance mindset. Indeed, a recent

national survey of provosts discussed in more detail below found that regional accreditors were the most significant drivers of assessment at colleges and universities, followed by program accreditors and then several measures of improvement. State and national calls were even further down the list (Kuh et al. 2014).

The seven regional organizations that accredit the overwhelming majority of not-for-profit education institutions have their own standards or criteria regarding student outcomes, but all require that institutions identify learning outcomes and undertake efforts to assess them. The Accrediting Commission for Community and Junior Colleges (ACCJC) of the Western Association of Schools and Colleges, the accreditor for associate's degree-granting institutions in California, Hawaii, and elsewhere, for example, has four accreditation standards, the first of which is built largely around student learning, its assessment, and efforts to improve it. Among the numerous relevant elements of the standard are:

- “The institution defines and assesses student learning outcomes for all instructional programs and student and learning support services” (ACCJC 2014, p. 2).
- “The institution uses assessment data and organizes its institutional processes to support student learning and student achievement” (ACCJC 2014, p. 2).
- “The institution broadly communicates the results of all of its assessment and evaluation activities so that the institution has a shared understanding of its strengths and weaknesses and sets appropriate priorities” (ACCJC 2014, p. 2).
- “The institution uses documented assessment of student learning and evaluation of student achievement to communicate matters of academic quality to appropriate constituencies, including current and prospective students and the public” (ACCJC 2014, p. 3).

Effective July 1, 2016, the New England Association of Schools and Colleges' (NEASC) Commission on Institutions of Higher Education (CIHE)—which accredits colleges and universities in the northeastern portion of the country—has nine standards, the eighth of which, Educational Effectiveness, is broadly outlined as:

The institution demonstrates its effectiveness by ensuring satisfactory levels of student achievement on mission-appropriate student outcomes. Based on verifiable information, the institution understands what its students have gained as a result of their education and has useful evidence about the success of its recent graduates. This information is used for planning and improvement, resource allocation, and to inform the public about the institution. Student achievement is at a level appropriate for the degree awarded. (CIHE 2016)

The subsections that follow include specific requirements around multiple forms of assessment of student learning and the use of the results for improvement.

Although most of the regional accreditors have updated their standards since its writing, or are currently in the process of doing so, Provezis's (2010) paper for the National Institute for Learning Outcomes Assessment (NILOA) remains one of the most useful overviews of accreditation practices and assessment. She concluded that each of the accreditors expected that outcomes be “defined, articulated, assessed, and used to guide institutional improvement” (p. 7). None of the accreditors mandated specific measures or approaches to be used, instead calling on institutions to use the methods most appropriate for their missions and characteristics. All

but one required that faculty be involved in assessment activities and even that one had an expectation that they would be. While there are some differences in the specifics of accreditors' efforts and approaches, recent evidence indicates that the outcomes are more similar than they are different (Gannon-Slater et al. 2014).

Provezis's (2010) report also highlighted the role that most regional accreditors play in providing resources and training to help their institutional members implement and improve assessment practices. Maki (2010) likewise noted "Creating additional resources, workshops, institutes, and guidelines, accreditors are pressing institutions to prioritize as well as mature the assessment process so that it leads to changes in pedagogy, curricular and instructional design, and educational practice" (p. 10). These broader efforts elucidate the oft-ignored dual nature of accreditors' roles. The accountability mandate that is enacted through accreditation process and has pushed institutions to undertake more learning outcomes assessment is joined by educational components that help foster better assessment practices on campuses. At the same time, the accreditors' roles in promoting assessment have been conflicted, and institutional response has been uneven. Their very efforts that have caused institutions to pay greater attention to and enact more forms of assessment have too often led to a compliance mentality, rather than a student-centered emphasis on improvement (Kuh et al. 2015).

Added to these institutional accreditors are program-specific accreditors that have varying degrees of influence over academic programs, departments, and schools. Fields such as nursing (Commission on Collegiate Nursing Education and National League for Nursing Accrediting Commission), business (Association to Advance Collegiate Schools of Business), education (Council for Accreditation of Educator Preparation), and pharmacy (American Council on Pharmaceutical Education) have strong accrediting bodies that consider student learning outcomes as part of their evaluations (Maki 2010; Palomba 2002). Maranville et al. (2012) highlighted the great disparities in approaches and requirements between fields. Legal education, they argued, was "ten to twenty-five years behind engineering and other professions" (p. 1027). Yet at the time they were writing, the American Bar Association's (ABA) Standards Review Committee was in the process of considering whether to require institutions to assess student learning outcomes as part of their accreditation process; in 2014, newly adopted standards included "A law school shall utilize both formative and summative assessment methods in its curriculum to measure and improve student learning and provide meaningful feedback to students" (ABA 2014). While the ABA may have come to learning assessment later than some accreditors, many fields do not have program accreditation at all and, as such, lack the prod toward outcomes that they can provide. Those that do are often more committed to and advanced in learning outcomes assessment practices than those that do not (Palomba and Banta 2001).

Higher Education Membership Associations

Accreditors are the most important membership organizations in fostering outcomes assessment, but institutions frequently belong to other mission-driven higher education associations. These associations, generally organized around institutional types

and shared interests, lack the rule of law or accountability roles yet can help push conversations, research, and efforts among their members while also helping to shape the national conversations and policies around learning outcomes. The aforementioned VSA, put forth by APLU and AASCU, is a prime example of the latter. The AACU's work on outcomes assessment includes the influential Valid Assessment of Learning in Undergraduate Education (VALUE), a campus-based rubric development project begun in 2007 as part of the association's broader Liberal Education and America's Promise (LEAP) initiative. Since the first iterations of VALUE rubrics were launched, thousands of institutions in the USA and beyond have used them to assess learning across 16 specific outcomes: civic engagement, creative thinking, critical thinking, ethical reasoning, foundations and skills for life-long learning, global learning, information literacy, inquiry and analysis, integrative and applied learning, intercultural knowledge and competence, oral communication, problem-solving, quantitative literacy, reading, teamwork, and written communication. The rubrics help individual instructors, campuses, and multi-institutional groups of faculty not only to understand what students have learned but to have conversations across disciplines about what they are expected to. In the most successful instances, the results have then been used to reconsider practices to improve learning. Moreover, VALUE and LEAP align with the Lumina Foundation's Degree Qualifications Profile (DQP), an effort to provide standard baselines for what students should know and be able to do to earn degrees at the associate's, bachelor's, and master's levels (Rhodes and Finley 2013; <https://www.aacu.org/value>). In conjunction with the Association of State Higher Education Executive Officers (SHEEO), AACU has launched the Multi-state Collaborative to Advance Learning Outcomes Assessment, involving more than 900 faculty members at 80 institutions in 13 states. The project—which utilizes the VALUE rubrics to assess authentic student work rather than an externally imposed task—is thought by some assessment advocates to be among the most promising of current efforts to understand and capture student learning (Berrett 2016). Numerous other membership associations, including the Council of Independent Colleges and the Association of Governing Boards of Universities and Colleges, are likewise working with their membership to promote assessment practices and use (Kinzie et al. 2015).

Disciplinary and professional associations that count individuals as members are very different than institutional associations and have not, historically, been as heavily involved in outcomes assessment (Hutchings 2011). Yet, in the current climate, they, too, have begun to engage with the issues, at times pushing for more attention and commitment to understanding what students learn in their disciplines and fields. Organizations such as the American Psychological Association have offered leadership in assessment, creating materials for members to use in considering and measuring outcomes. Professional associations, such as the American College Personnel Association, the NASPA-Student Affairs Administrators in Higher Education, and the National Academic Advising Association, have all emphasized assessment for student affairs professionals (Maki 2010). At times, though, these efforts have been contested, especially in humanities associations. When, for example, then Modern Language Association (MLA) President Gerry

Graff (2008) announced that he was “a believer in the potential of learning outcomes assessment” (p. 3), many in his discipline were surprised, and some were dismayed (Bennett and Brady 2012; Feal et al. 2011). Yet, the MLA was already involved in learning outcomes through its Association of Departments of English and Association of Departments of Foreign Languages (Feal et al. 2011). Despite continuing concerns about whether assessment practices are able to capture the true products of education in the humanities, in 2010, all but 2% of English departments surveyed reported either having or planning an assessment program (Heiland and Rosenthal 2011). The American Historical Association (AHA) has likewise become more involved in assessment in the years since the Spellings Commission, with the understandings that disciplinary knowledge matters and that external groups might impose measures and assessments if the disciplines themselves do not (Hyde 2014). Since 2012, it has undertaken the AHA History Tuning Project as part of the larger Lumina-sponsored Tuning USA efforts building on the Bologna Process efforts in the US context. Historians at more than 120 colleges and universities have worked to define the knowledge, skills, and abilities that should be expected outcomes of a history education. Although tuning is not itself an assessment, it places faculty at the center of a larger process that sets the stage for assessment of these expected outcomes (McInerney 2016). While substantial differences remain between the efforts and commitments of different associations—and of their membership—the recognized importance of disciplinary knowledge has drawn many more into serious consideration of leaning outcomes.

Foundations

A final category of external nongovernmental actors requires some consideration as its constituents help provide significant resources for, and help shape the conversations around, learning outcomes assessment: philanthropic foundations. The Lumina Foundation’s DQP and Tuning USA projects have already been mentioned, but they are just part of the foundation’s broader efforts to promote understanding of what students learn and to create and disseminate resources to assess and align higher education outcomes. The foundation has convened faculty, fostered the creation of a library of assignments aligned with proficiencies in the DQP, supported the development of the VSA, and otherwise funded significant work on learning outcomes assessment. Much of it has been done in partnership with other organizations, including NILOA (www.learningoutcomesassessment.org). Founded in 2008 with support from Lumina, the Carnegie Corporation of New York, and the Teagle Foundation, NILOA is the leading institute dedicated to promoting research on, use of, and best practices in learning outcomes assessment. Through its various funded efforts, it has worked to shift the framing of assessment from an accountability paradigm to an improvement paradigm. Teagle has likewise been influential through its support of the Wabash National Study, Outcomes and Assessment initiatives, and efforts to promote institutions to work independently and together to improve assessment practices (Kinzie et al. 2015). Perhaps second only to accreditors, these and other foundations have driven learning outcomes conversations, although their efforts have not always been welcomed among the faculty.

2.5 Assessment on the Ground

While the governmental and extra-governmental actors are integral to fostering conversations about and pushing efforts for learning outcomes assessment, college and university campuses are where these efforts can come to fruition. Indeed, more institutions report engaging in assessment-related activities than even a few years ago. At the same time, the reaction to and experience of these activities remains problematic. Many in US tertiary education continue to view assessment as part of a larger accountability mandate rather than one that is ideally aimed at improving learning. Faculty, especially but not just in the humanities, question the value of assessment activities, raise concerns about the additional work that some plans require, and worry about both institutional autonomy and professorial academic freedom. Perhaps most pressingly, even when well undertaken with goals of improving learning, it remains difficult for institutions to complete the assessment cycle by undertaking meaningful reforms at the department, college, and institutional levels.

Internal Actors

The NILOA senior staff recently articulated the different roles that institutional constituencies can and should play in making assessment efforts consequential. Ewell and Ikenberry (2015) identified the importance of governing board and administrative leadership in assessment efforts. The former set broad guidelines ensure that the academic officers have the tools needed to undertake assessment and focus on the strategic use of results. Academic administrators—presidents, provosts, deans, and department chairs—act at different levels within colleges and universities but are all tasked with providing leadership, if in different ways. Presidents set strategic plans, provide oversight of accreditation activities, and oversee the provosts, who “have the central leadership role in assessment at any college or university” (p. 127). Provosts, or chief academic officers, oversee program review and faculty reward structures and are responsible for engaging with deans and other campus constituents about the practice and outcomes of assessment. Deans and department chairs undertake similar roles, if on more localized levels and in closer contact to actual assessment practices.

Cain and Hutchings (2015) pointed to the central role that academics and professionals closest to students—librarians, student affairs professionals, and most significantly faculty—play in assessment practices. With their historic control over the college curriculum and the current recognition that effective assessment efforts are tied closely to the work that is already being done in courses and through related academic experiences, faculty are perhaps the most important stakeholders in articulating what students should know, documenting their knowledge and skills, and undertaking efforts for improvement. The very attributes and roles that place faculty at the center of this assessment work, though, can also cause them to resist efforts that are viewed as merely responding to external accountability mandates, that seek to track student learning in crude rather than nuanced ways, or that threaten their authority or portend an evaluation scheme. As Cain and Hutchins contended, students too are vital participants in the process, not merely subjects upon whom assessment can be performed.

Surveys of Current Practices

Some of the best evidence of what is actually happening on college campuses comes from NILOA's 2013 national surveys of provosts, a follow-up to similar 2009 survey. (At the time of this writing, a third iteration is in the field.) With responses from provosts or designees from more than 1200 accredited 2- and 4-year colleges, the 2013 survey found increasing attention paid to learning outcomes and their assessment, though continuing difficulties in using the results for improvement. In 2013, 84% of surveyed institutions reported institution-level learning outcomes statements, compared with 74% just 4 years earlier. These statements, which are crucial building blocks to actual assessment, were in full alignment with programmatic learning statements more than 40% of the time, nearly 50% for all but doctoral institutions. As was the case in the earlier survey, regional and programmatic accreditations were the primary drivers of assessment activities. Importantly though, many also indicated that both institutional commitment to improve and faculty and staff interest in student learning were significant factors, as well (Kuh et al. 2014).

Among the most promising findings for advocates of assessment is the increased use of multiple measures to assess student learning, including the increased use of authentic measures that build on students' classroom activities. On average, institutions reported using five different measures, up from 4 years earlier. More than 80% of respondents indicated that their schools used national student surveys, while nearly 70% indicated the use of rubrics and classroom-based performance assessments. Almost all types of assessment tools were being used more frequently than in 2009, but the largest increases in use were of rubrics, portfolios, and external performance assessments such as internships or service learning. As in the earlier survey, assessment results were most often used for accreditation and accountability purposes, but the 2013 iteration revealed that more institutions were using them for program review, curriculum modification, learning goals revision, and institutional improvement. More than 70% of provosts indicated "quite a bit" or "very much" organizational support for assessment, and many indicated the importance of institutional policies, faculty engagement, assessment committees, and institutional research in pursuing their assessment efforts. Yet challenges remain. Among the most pressing concerns noted was the need for greater faculty engagement, greater faculty use of results, and better professional development for faculty. Provosts also indicated a need for more useful and productive assessment, concerns about maintaining momentum, and difficulties in creating a culture of assessment where it is viewed as a process integral to the institutions' functioning, rather than an effort to appease accountability demands (Kuh et al. 2014).

The NILOA surveys demonstrated that the assessment practices differ across institutional types in the highly diversified US system. Broadly speaking, the more selective an institution, the less developed its assessment practices. Doctoral universities were more likely to use national surveys and less likely to use most other measures than any other institutional type. They were also less likely to use results of assessment activities for program review, curricular change, learning goal revision, strategic planning, or other activities other than accreditation (Kuh et al. 2014). The Association of American Universities (AAU 2013), a membership association

of 60 US and 2 Canadian research universities, proximally surveyed its membership on assessment practices. Based on responses from 37 institutions, the association argued that the perception that its membership neglected undergraduate education was misguided and pointed to myriad assessment practices, which differed between program and institutional levels. At the same time, it highlighted that AAU schools often found national standardized tests such as the Collegiate Learning Assessment or the ETS Proficiency Profile to be ill suited for their purposes and misaligned with their intended outcomes. They were more likely to use discipline-specific and faculty-driven tools, especially at the program level. At the same time, pushed in part by accreditation demands, assessment efforts were becoming more centralized at many of the responding institutions.

AACU surveyed 325 member provosts or designees across institutional types in 2015 and supplemented the findings with a small set of interviews. The results, conveyed in three reports in late 2015 and early 2016, included that 85% of responding colleges had common intended learning outcomes for all undergraduates compared with 78% 7 years earlier. Moreover, there was broad consensus across the skills and knowledge areas covered by the outcomes. More than three quarters indicated clear learning outcomes for their general education programs, and more than two-thirds indicated both that they assessed the achievement of those outcomes and that they were clearly linked to students' requirements. Similar to the NILOA surveys, institutions reported using a variety of methods to assess learning, although the specific favored measures were somewhat different. Most common for assessing general education were locally created rubrics, followed by capstone projects, student surveys, and common assignments across courses. Standardized national tests of skills and knowledge were the least used. Many institutions used e-portfolios, though few required them of all students. While many of the findings were positive, the reports also revealed that institutions were not adept in conveying the intended outcomes to students. Additionally, while most institutions indicated goals to reduce gaps in graduation rates between racial and ethnic groups, few had plans to address learning outcomes gaps between such groups; even fewer did for different socioeconomic groups or for students with different parental educational attainments (Hart Research Associates 2015, 2016a, b).

Several more recent surveys have touched on related issues. An Inside Higher Ed survey revealed that many faculties remained doubtful that assessment improves student learning and most viewed it as designed to meet external reporting needs (Jaschik and Lederman 2016). The consulting firm Eduventures's survey of more than 200 higher education leaders identified disagreements and confusion over who on campus was responsible for student outcomes and whether those outcomes were related to learning or otherwise (Wiley 2016). Preliminary results from Taskstream's recent study built on focus groups, interviews, and survey responses with institutional leaders, assessment professionals, and faculty revealed that only 25% of respondents indicated that their institution's assessment efforts were "mature." Almost half reported that the term "outcome" lacked a clear definition (Curtis and Gulliford 2016).

Taken together, these surveys show the conflicted space of learning outcomes assessment in US tertiary education. More than a decade since the Spellings Commission catalyzed increased interest and effort, assessment practices have yet

to become fully integrated into colleges and universities. Faculties remain skeptical of the value of assessment and dubious of its external drivers. There is widespread confusion about the terminology that is used and the jargon that can accompany expanding assessment programs. At the same time, more assessment of learning outcomes is taking place, and more of it relies on homegrown measures that tie directly to the work that students and faculty already do. Institutions are increasingly linking those efforts to improvement, although the extent of actual improvement is far less clear.

2.6 Conclusion

Assessing student learning remains an important and difficult challenge—Marinara et al. (2004) likened it to capturing a “loose baggy monster”—and using the results of that assessment to improve education is even more difficult. While the best evidence indicates that there is still significant work to be done, there is no doubt that US colleges and universities are more involved with assessing student learning than even a few years ago. The national mandate from accreditors, state-level accountability efforts, and infrastructure built through disciplinary and membership associations have all contributed to increased attention and refined practices. Perhaps most promisingly, institutions are increasingly turning to the work that faculty and students are already engaged with to understand what students learn and are able to do. These efforts portend increased awareness of the value of embedding assessment into ongoing faculty-driven practices in cooperation with those most closely involved with student learning.

Despite some encouraging signs, significant challenges remain if assessment is to be fully embraced and capitalized upon in US higher education. Accountability demands are legitimate, but a primary concern remains shifting the framing of assessment from compliance to improvement. When driven by external mandates, assessment results are less likely to be used to reconsider and reform practices. Flipping the approach to focus on improvement and then using the by-products to satisfy accountability requirements offer more promise (Kuh et al. 2015). Closely related is the need to generate further buy-in, especially among the students and faculty at the center of the process. Students are too often left out altogether, and the very real disciplinary and pedagogical expertise that faculty possess is too often neglected (Cain and Hutchings 2015). Moreover, amid broader changes in American higher education, faculties are rightfully worried about encroachments into their classrooms and curricular prerogatives, as well as threats posed by increasingly business-oriented decision-making (see also Coates, Chap. 1 in this volume). While assessment efforts were at one time linked to business practices, they do not need to be and, when well constructed, do not need to threaten academic freedom (Cain 2014). If such efforts are documented through rigorous research that shows that the human and financial resources expended on it are justified by improved student learning while simultaneously meeting accountability requirements, learning outcomes assessment will meet perhaps its ultimate challenge.

References

- Accrediting Commission for Community and Junior Colleges-Western Association of Schools and Colleges. (2014). *Accreditation standards*. Retrieved from http://www.accjc.org/wp-content/uploads/2014/07/Accreditation_Standards_Adopted_June_2014.pdf
- Allen, C. (2016, July). *Alverno College: Lessons from an assessment pioneer* (NILOA examples of good assessment practice). Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. Retrieved from <http://www.learningoutcomesassessment.org/AlvernoCaseStudy.html>
- American Association of State Colleges and Universities. (2011). *A guide to major U.S. college completion initiatives. A policy matters report*. Washington, DC: American Association of State Colleges and Universities. Retrieved from <http://www.aascu.org/policy/publications/policymatters/2011/collegecompletion.pdf>
- American Bar Association Section of Legal Education and Admissions to the Bar. (2014). *Revised standards for approval of law schools*. Retrieved from http://www.americanbar.org/content/dam/aba/administrative/legal_education_and_admissions_to_the_bar/council_reports_and_resolutions/201406_revised_standards_clean_copy.authcheckdam.pdf
- Association of American Colleges & Universities. (1985). *Integrity in the college curriculum: A report to the academic community*. Washington, DC: AAC&U.
- Association of American Colleges & Universities. (2013). *AAC&U responds to President Obama's proposed college ratings system*. Retrieved from <https://www.aacu.org/about/statements/2013/ratings>
- Association of American Universities. (2013). *AAU survey of undergraduate student objectives and assessment*. Retrieved from <http://www.aau.edu/WorkArea/DownloadAsset.aspx?id=14849>
- Bastedo, M. N., & Bowman, N. A. (2010). U.S. news & world report college rankings: Modeling institutional effects on organizational reputation. *American Journal of Education*, 116, 163–183.
- Bastedo, M. N., & Bowman, N. A. (2011). College rankings as an interorganizational dependency: Establishing the foundation for strategic and institutional accounts. *Research in Higher Education*, 52, 3–23.
- Bastedo, M. N., & Jaquette, O. (2011). Running in place: Low-income students and the dynamics of higher education stratification. *Educational Evaluation and Policy Analysis*, 33, 318–339.
- Bennett, M., & Brady, J. (2012). A radical critique of the learning outcomes assessment movement. *Radical Teacher*, 94, 34–44.
- Berrett, D. (2016, October 16). The next great hope for measuring learning. *The Chronicle of Higher Education*. Retrieved from http://www.chronicle.com/article/The-Next-Great-Hope-for-1238075?key=BFxeuhOx7Lbr_3ptAXUCm5IIvJor4pwKHkHrDm8SaDLaNDCvMwyM9HwIsJcurlnNVzF5c0o2U2NLLW5Wa1RoZ2IIUzUzMU55SFQtQVRKY1E4aDJocjZIZUtuZw
- Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate*. Princeton: Carnegie Foundation for the Advancement of Teaching.
- Cain, T. R. (2014, November). *Assessment and academic freedom: In concert, not conflict* (NILOA Occasional Paper No. 22). Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. Retrieved from <http://learningoutcomesassessment.org/occasionalpapertwentytwo.html>
- Cain, T. R., & Hutching, P. (2015). Faculty and students: Assessment at the intersection of teaching and learning. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie (Eds.), *Using evidence of student learning to improve higher education* (pp. 95–116). San Francisco: Jossey-Bass.
- Chronicle of Higher Education Almanac. (2016). Retrieved from http://www.chronicle.com/interactives/almanac-2016#id=2_101
- Clark, B. R. (1979). The many pathways of academic coordination. *Higher Education*, 8, 251–267.
- Commission on Institutions of Higher Education. (2016). *Standards for accreditation*. Burlington: Commission on Institutions of Higher Education. Retrieved from https://cihe.neasc.org/sites/cihe.neasc.org/files/downloads/Standards/Standards_for_Accreditation.pdf

- Curtis, M., & Gulliford, M. (2016, October 26). *The “state of affairs” of assessment in higher education: Preliminary survey findings [Webinar]*. New York: Taskstream.
- Doyle, W. R. (2010). Changes in institutional aid, 1992–2003: The evolving role of merit aid. *Research in Higher Education*, 51, 789–810.
- Ewell, P. T. (2002). An emerging scholarship: A brief history of assessment. In T. W. Banta & Associates (Ed.), *Building a scholarship of assessment* (pp. 3–25). San Francisco: Jossey-Bass.
- Ewell, P. T., & Ikenberry, S. O. (2015). Leadership in making assessment matter. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie (Eds.), *Using evidence of student learning to improve higher education* (pp. 117–145). San Francisco: Jossey-Bass.
- Feal, R. G., Laurence, D., & Olsen, S. (2011). Where has assessment been in the modern language association? A disciplinary perspective. In D. Heiland & L. J. Rosenthal, with C. Ching, (Eds.), *Literary study, measurement, and the sublime: Disciplinary assessment* (pp. 9–24). New York: Teagle Foundation.
- Field, K. (2013, August 22). Obama plan to tie student aid to college ratings draws mixed reviews. *The Chronicle of Higher Education*. Retrieved from <http://www.chronicle.com/article/obama-plan-to-tie-student-aid/141229>
- Finkelstein, M. J., Conley, V. M., & Schuster, J. H. (2016). *The faculty factor: Reassessing the American academy in a turbulent era*. Baltimore: Johns Hopkins University Press.
- Gannon-Slater, N., Ikenberry, S., Jankowski, N., & Kuh, G. (2014, March). *Institutional assessment practices across accreditation regions*. Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. Retrieved from <http://www.learningoutcomeassessment.org/documents/Accreditation%20report.pdf>
- Gaston, P. L. (2014). *Higher education accreditation: How it’s changing, why it must*. Sterling: Stylus.
- Graff, G. (2008). Assessment changes everything. *MLA Newsletter*, 40(1), 3–4.
- Hart Research Associates. (2015). *Bringing equity and quality learning together: Institutional priorities for tracking and advancing underserved students’ success*. Washington, DC: AACU. Retrieved from <http://www.aacu.org/sites/default/files/files/LEAP/2015AACUEquityReport.pdf>
- Hart Research Associates. (2016a). *Recent trends in general education design, learning outcomes, and teaching approaches*. Washington, DC: AACU. Retrieved from http://www.aacu.org/sites/default/files/files/LEAP/2015_Survey_Report2_GEtrends.pdf
- Hart Research Associates. (2016b). *Trends in learning outcomes assessment*. Washington, DC: AACU. Retrieved from http://www.aacu.org/sites/default/files/files/LEAP/2015_Survey_Report3.pdf
- Hearn, J. C. (2015). *Outcomes-based funding in historical and comparative contexts*. A Lumina issue paper prepared for HCM strategists. Indianapolis: Lumina Foundation. Retrieved from <https://www.luminafoundation.org/files/resources/hearn-obf-full.pdf>
- Hearn, J. C., & Deupree, M. M. (2013). Here today, gone tomorrow? The increasingly contingent faculty workforce. A report prepared for the *Advancing Higher Education* publication series of the TIAA-CREF Institute, March. New York: TIAA-CREF Institute. Retrieved from http://www.tiaa-crefinstitute.org/institute/research/advancing_higher_education/contingent-faculty0313.html.
- Hearn, J. C., & Holdsworth, J. M. (2002). Influences of state-level policies and practices on college students’ learning. *Peabody Journal of Education*, 73(3), 6–39.
- Hearn, J. C., & Rosinger, K. O. (2014). Socioeconomic diversity in selective private colleges: An organizational analysis. *Review of Higher Education*, 38, 71–104.
- Heiland, D., & Rosenthal, L. J. (2011). Introduction. In D. Heiland & L. J. Rosenthal, with C. Ching, (Eds.), *Literary study, measurement, and the sublime: Disciplinary assessment* (pp. 9–24). New York: Teagle Foundation.
- Hillman, N. (2016). *Why performance-based college funding doesn’t work*. Report for The Century Foundation. Retrieved from <https://tcf.org/content/report/why-performance-based-college-funding-doesnt-work/>

- Hovey, H. A. (1999). *State spending for higher education in the next decade: The battle to sustain current support* (National Center Report 99-3). Washington, DC: National Center for Public Policy and Higher Education.
- Hutchings, P. (2011). From department to disciplinary engagement. *Change*, 41(3), 26–33.
- Hyde, A. (2014). Tuning and teaching history as an ethical way of being in the world. *AHA Today: A Blog of the American Historical Association*. Retrieved from <http://blog.historians.org/2014/07/tuning-teaching-history-ethical-way-world/>
- Ikenberry, S. O., & Kuh, G. D. (2015). From compliance to ownership: Why and how colleges and universities assess student learning. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie (Eds.), *Using evidence of student learning to improve higher education* (pp. 1–23). San Francisco: Jossey-Bass.
- Jaschik, S., & Lederman, D. (2016). *The 2016 Inside Higher Ed survey of faculty attitudes on technology*. Washington, DC: Inside Higher Ed & Gallup. Retrieved from <https://www.inside-highered.com/system/files/media/2016%20IHE%20Faculty%20Tech%20Survey.pdf>
- Kinzie, J., Ikenberry, S. O., & Ewell, P. T. (2015). The bigger picture: Student learning outcomes assessment and external entities. In G. D. Kuh, S. O. Ikenberry, N. A. Jankowski, T. R. Cain, P. T. Ewell, P. Hutchings, & J. Kinzie (Eds.), *Using evidence of student learning to improve higher education* (pp. 160–180). San Francisco: Jossey-Bass.
- Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). *Knowing what students know and can do: The current state of learning outcomes assessment in U.S. Colleges and Universities*. Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.
- Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, T. R., Ewell, P. T., Hutchings, P., & Kinzie, J. (2015). *Using evidence of student learning to improve higher education*. San Francisco: Jossey-Bass.
- Maki, P. L. (2010). *Assessing for learning: Building a sustainable commitment across the institution* (2nd ed.). Sterling: Stylus.
- Maranville, D., O’Neill, K., & Plumb, C. (2012). Lessons for legal education from the engineering profession’s experience with outcomes-based accreditation. *William Mitchell Law Review*, 38, 1017–1093.
- Marinara, M., Vajravelu, K., & Young, D. L. (2004). Making sense of the “loose baggy monster”: Assessing learning in a general education program is a whale of a task. *Journal of General Education*, 53, 1–19.
- McInerney, D. J. (2016). The American historical associations tuning project: An introduction. *The History Teacher*, 49, 491–501.
- McLendon, M. K., & Hearn, J. C. (2009). Viewing recent U.S. governance reform whole: “Decentralization” in a distinctive context. In J. Huisman (Ed.), *International perspectives on the governance of higher education: Alternative frameworks for coordination* (pp. 161–181). New York: Routledge.
- National Center for Higher Education Management Systems. (2016). *College-going rates of high school graduates-directly from high school-2010*. Retrieved from <http://www.higheredinfo.org/dbrowser/index.php?measure=32>
- National Governors’ Association. (1986). *Time for results: The governors’ 1991 report on education*. Washington, DC: National Governors’ Association.
- National Institute of Education, The (NIE). (1984). *The involvement in learning: Report of the commission on quality in American higher education*. Washington, DC: U.S. Government Printing Office.
- Newman, F. (1987). *Choosing quality: Reducing conflict between the state and the university*. Denver: Education Commission of the States.
- Palomba, C. A. (2002). Scholarly assessment of student learning in the major and general education. In T. W. Banta (Ed.), *Building a scholarship of assessment* (pp. 201–222). San Francisco: Jossey-Bass.

- Palomba, C. A., & Banta, T. W. (2001). *Assessing student competence in accredited disciplines: Pioneering approaches to assessment in higher education*. Sterling: Stylus.
- Powell, C. (2013). Accreditation, assessment, and compliance: Addressing the cyclical challenges of public confidence in American higher education. *Journal of Assessment and Institutional Effectiveness*, 3, 54–74.
- Provezis, S. (2010). *Regional accreditation and student learning outcomes: Mapping the territory*. Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.
- Public Agenda. (2016). *Public opinion on higher education: Our new survey suggests public confidence in higher education is waning*. Report released September 12, 2016. Retrieved from <http://www.publicagenda.org/pages/public-opinion-higher-education-2016>
- Rhodes, T. L., & Finley, A. (2013). *Using the VALUE rubrics for improvement of learning and authentic assessment*. Washington, DC: Association of American Colleges & Universities. Retrieved from <https://www.aacu.org/value>
- Schmidtlein, F. A., & Berdahl, R. O. (2011). Autonomy and accountability: Who controls academe? In P. G. Altbach, P. J. Gumpert, & R. O. Berdahl (Eds.), *American higher education in the 21st century: Social, political, and economic challenges* (pp. 69–87). Baltimore: Johns Hopkins University Press.
- Schuster, J., & Finkelstein, M. (2006). *The American faculty: The restructuring of academic work and careers*. Baltimore: Johns Hopkins University Press.
- Shavelson, R. J. (2007). *A brief history of student learning assessment: How we go where we are and a proposal for where to go next*. Washington, DC: AACU. Retrieved from http://cae.org/images/uploads/pdf/19_A_Brief_History_of_Student_Learning_How_we_Got_Where_We_Are_and_a_Proposal_for_Where_to_Go_Next.PDF
- Sims, S. J. (1992). *Student outcomes assessment: A historical review and guide to program development*. New York: Greenwood Press.
- Smith, A. A. (2015). Uniting to regulate for-profits. *Inside Higher Education*. Retrieved from <https://www.insidehighered.com/news/2015/10/12/federal-agencies-join-forces-regulate-profit-colleges>.
- Trow, M. A. (1984). The analysis of status. In B. R. Clark (Ed.), *Perspectives on higher education: Eight disciplinary and comparative views* (pp. 132–164). Berkeley: University of California Press.
- Trow, M. (2007). Reflections on the transition from elite to mass to universal access: Forms and phases of higher education in modern societies since WWII. In J. F. Forest & P. G. Altbach (Eds.), *International handbook of higher education* (pp. 243–280). Dordrecht: Springer.
- U.S. Census Bureau. (2016). *Current population survey*. Available at: <http://www.census.gov/programs-surveys/cps.html>
- U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Retrieved from <http://www2.ed.gov/about/bdscomm/list/hiedfuture/reports/final-report.pdf>
- U.S. Department of Education. (2014). *Digest of education statistics, 2014*. Retrieved from <http://nces.ed.gov/pubs2016/2016006.pdf>
- Wiley, J. (2016). *Hitting an unclear target: The impact of ambiguous 'student outcomes' on technology*. Retrieved from <http://www.eduventures.com/2016/10/hitting-an-unclear-target-the-impact-of-ambiguous-student-outcomes-on-technology/>

Chapter 3

Information Management Versus Knowledge Building: Implications for Learning and Assessment in Higher Education



Patricia A. Alexander

Abstract The goal of this paper is to consider two distinct orientations toward learning within the context of twenty-first-century higher education that have implications for assessment of outcomes internationally—information management and knowledge building. These two orientations are compared and contrasted along various dimensions, and potential contributors to the pervasiveness of the information management profile within the current generation of undergraduates are explored. With this background established, pertinent steps toward fostering more effective information management and enhancing knowledge building in higher education contexts are shared with specific attention to the role of assessment practices.

3.1 Introduction

All truths are easy to understand once they are discovered; the point is to discover them.
(Galileo Galilei)

The claim I proffer in this paper is that there are two general perspectives on learning within the context of higher education—information management and knowledge building—and that these perspectives have significant implications for learning and assessment at the tertiary level. I opened this discussion with the quote by Galileo precisely because I only came to this “discovery” about information management and knowledge building quite recently, after decades of research on learning and academic development (Alexander et al. 1995; Alexander and Murphy 1998) and cognitive assessment (Alexander et al. 2015, 2016). Fundamentally, what I came to realize was that twentieth-century beliefs about the nature and process of

P. A. Alexander (✉)
University of Maryland, College Park, MD, USA
e-mail: palexand@umd.edu

learning, which still prevail in higher education, may be blinding educators to the academic reality that exists for many twenty-first-century college students.

The resulting disparity between these twentieth-century “learning” beliefs and educational realities for twenty-first-century students can have negative consequences for teaching and assessment practices within higher education (see also Coates, Chap. 1, in this volume). Moreover, unless educators and university students adopt alternative mind-sets about what it means to teach and what it means to learn, this disparity will persist and potentially exacerbate. I appreciate that theories and models addressing different goals, orientations, or approaches to learning within higher education and their potential consequences for teaching and assessment are by no means new (Biggs and Collis 1982; Entwistle and Entwistle 2003; Meece et al. 1988). However, I would argue that the two perspectives introduced in this paper are broader in conceptualization and can encompass aspects of those differing goals, orientations, and approaches described in the literature.

Therefore, my goal for this paper is to bring the awareness about information management and knowledge building to the global higher education community responsible for teaching and assessing the current generation of college students. I will also consider the reasons for this shift in how students perceive the purpose and nature of higher education. Further, within this volume, which is devoted to looking at issues of learning and assessment within higher education from an international perspective, it is important to establish that the orientation toward information management over knowledge building is not solely an American phenomenon. The very circumstances that give rise to information management in US classrooms exist in postindustrial societies globally, as I will endeavor to show (e.g., Shavelson et al. in this volume). In concluding the paper, I will consider potential actions that can be taken to reframe teaching and assessment in higher education and to readjust faculty and students’ perspectives toward learning in university classrooms.

I recognize that the discussion within this paper is rather theoretical and somewhat speculative in nature. Yet, even though the conceptualization of information management and knowledge building I present began as a theoretical insight, it can be empirically substantiated by studies from diverse literatures. Those literatures include investigations in learning and strategic behavior (Entwistle and Peterson 2004; Marton and Säljö 1976), epistemic cognition (e.g., Franco et al. 2012; Mason et al. 2011), multiple source use (e.g., Braasch et al. 2012; List et al. 2015; List and Alexander 2017), technological use patterns (e.g., Rideout et al. 2013; Singer and Alexander 2017; Wood et al. 2012), expertise development (Alexander 1997; Nandogopal and Ericsson 2012), rational thinking (e.g., Evans and Stanovich 2013; Stanovich et al. 2011), and relational reasoning (Alexander et al. 2015; Dumas et al. 2014). Throughout this paper, those literatures will serve as the empirical backbone for the claims and recommendations forwarded.

3.2 Contrasting Information Management and Knowledge Building

3.2.1 Information Management

Broadly speaking, information management can be defined as the manipulation of data from multiple sources and the organization, regulation, and communication of that data to multiple audiences in multiple forms and for multiple purposes (Alexander 2015a, b). What is particularly salient about information management is that this process is generally undertaken to accomplish some immediate and short-term task, with no intention of long-term retention of the resulting information. Provided that information presented or located appears on the surface to satisfy task parameters, there is no reason to delve into its veracity or to investigate alternative or confirming sources. Moreover, the value or utility of any information assembled through this process dissipates as soon as the task has been satisfied.

So described, information management resembles what is referred to as System 1 thinking (Stanovich et al. 2011). System 1 “thinking” has been characterized as cognitive processing that is quite shallow and reactive, rather than reflective or evidence-based in nature. Those students engaged in information management are often working, sometimes diligently, to keep their heads above the proverbial information floodwaters. As such, these students may have neither the time nor the inclination to delve deeper into the educational experience or to reflect critically on the information before them (e.g., Oser et al., Chap. 7; Shavelson et al., Chap. 10 in this volume). Consequently, the academic engagement of students adopting an information management perspective frequently remains shallow and reactive.

What further characterizes information management is that cognitive activities are often undertaken in the service of some “given” or externally directed task (see Table 3.1). Further, in that these “given” tasks frequently have little personal relevance to them, college students elect not to dive any deeper into the informational waters than necessary because they perceive no value in doing so—nothing to be gained by exerting additional effort (Bok 2009; Zusho et al. 2003). It is this reality that is particularly relevant to this volume on the assessment of learning in higher education (e.g., Brückner and Zlatkin-Troitschankaia, Chap. 6, in this volume).

Table 3.1 Comparing information management and knowledge building

Information management	Points of comparison	Knowledge building
Externally directed	Purpose	Self-formulated
Likely constrained	Effort	Potentially expansive
Relatively brief	Residual effects	More enduring
Mainly other-determined	Evaluation source	Largely self-determined
Pervasive	Frequency	More selective

In essence, students appear to understand their “job” to be completing their various academic tasks, and if that “job” can be satisfied with shallow and brief processing, then why do more? College students tick off their academic “to-do” list—complete assignments, study for tests, participate in classroom activities—because it is required of them in order to receive the grades or scores they want. Whether the knowledge associated with those duties is retained beyond class examinations or course completion or whether the content memorized is even accurate or well justified has little bearing on their thinking and actions (e.g., Hofer 2000). In fact, such epistemic reflection could complicate the mission of getting the academic work done efficiently. As compliant and grade-conscious students, undergraduates may surrender their personal epistemic authority by simply accepting what they read or hear as “truth” not caring enough to probe or challenge the ideas conveyed in those educational contexts as long as their responses receive full credit from those with the power to grade them.

I certainly do not want to leave the impression that college students, who seemingly lack the drive or ability to be deep thinkers, are either the sole or even primary source of this resistance to reflective and analytic thinking. Quite to the contrary, the argument can be readily made that these are students that the educational system has helped to create, as I will elaborate when I discuss the varied contributors to the disparity between information management and knowledge building. Nor do I wish to leave the impression that information management is a villain in higher education learning and assessment. Effective information management is a critical competency for any twenty-first-century student. Further, as a task-directed operation, effective information management seemingly entails valuable and foundational cognitive skills, such as:

- Ascertaining the nature of a given task
- Focusing attention and perception on task components
- Distinguishing relevant from irrelevant content
- Retaining information as long as needed
- Applying acquired information to the task at hand

Rather, the core question to be weighed here is whether twenty-first-century college students are *effective* managers of information; that is, do they manifest the aforementioned characteristics in their efforts to deal with the continual onslaught of informational? Further, do students ply these foundational skills intelligently, depending on the situation or context? In effect, do students possess the conditional knowledge to know *when* information management is most advantageous to them, as opposed to when more critical-analytic engagement would be of greater value?

Given the nature of the information-saturated world in which twenty-first-century students live and learn, it is understandable that students are, by necessity, managers of information. In addition, cast as “digital natives” (Prensky 2001), many of today’s college students seemingly operate under the mistaken perception that they are, in fact, highly skilled at these managerial duties (Selwyn 2003; Singer and Alexander 2017). Multitasking and pervasiveness of social media in their lives are just two of the contributors to such self-perceptions. However, there is ample evidence that this generation’s facility with and sometimes “addiction” to their smartphones, computers, and all things technological may have long-term consequences and serious

effects on the nature of mental processing in which they engage and the attention and reflection they allocate (Hembrooke and Gay 2003; Singer and Alexander 2017; Richtel 2010; Rosen et al. 2013).

3.2.2 Knowledge Building

In contrast to information management, knowledge building typically pertains to the analysis and processing of data with the intent of testing its veracity and utility and with the implicit or explicit goal of retaining it in memory for use at a future time (Table 3.1). By definition, knowledge building reflects learner intentions (Kulikowich and Alexander 2010) and a recognition that cognitive processing will need to be extended and deep enough to ensure that pertinent information leaves an enduring mark on memory (Dinsmore and Alexander 2012). Thus, knowledge building as a mental activity presumably requires:

- Establishing one's intentions
- Identifying importance or salience of the content
- Gathering sufficient evidence based on disciplinary or domain standards
- Engaging in critical or reflective analysis
- Judging the veracity or credibility of the information read or heard

Thus, knowledge building would appear to require more cognitive effort and different mental processes than information management (Alexander 2015b). Moreover, in those instances when students have no personal investment in the content of the classes they take, and if the forms of assessment within that context do not prompt students to delve deeply in the information presented in class or require them to justify the conclusions they reach, then information management may well prove sufficient for their academic aims. I have witnessed firsthand that many of my classes are populated with good students who are diligently performing their educational roles as they understand them to be. Regrettably, in so doing, these students may remain blissfully unaware that whatever potential for knowledge building exists within them and whatever rudimentary habits of mind associated with knowledge building lie dormant are slowly becoming atrophied. In effect, after so many years within the educational system, these undergraduates' efforts to be good students have made it increasingly more difficult for them to be "good learners" (Alexander 2015a).

3.3 Why Does This Situation Exist?

As educational researchers concerned with learning and assessment within higher education, we are dedicated not only to uncovering the patterns or trends that exist but also to exploring contributory factors that help explain the emergence of these

patterns or trends. I have already noted several widespread and powerful forces at work in the lives of twenty-first-century college students that appear implicated in their tendency to manage information more than build knowledge. The most notable of those forces is the sheer volume of information that invades contemporary lives—to say nothing about the speed at which that information is delivered, its diverse representations (e.g., graphic, numeric, or linguistic), and its varied means of conveyance (e.g., visual or auditory). This informational influx and rapidity, combined with competing events in college students' lives (e.g., school, work, or social), may help to underscore these individuals' sometimes fragmented and distributed attention (e.g., Foehr 2006; Rosen et al. 2013). College students may, by necessity, become multitaskers, but that does not mean such multitasking serves them well in or out of the classroom (Ophir et al. 2009; Richtel 2010).

In addition to the effects of contending with information saturation, technology has a role to play in students' tendency to engage in information management. There is ample evidence that traditional print modality is losing ground to digital, multimedia technologies (e.g., Rideout et al. 2013). Today's undergraduates have always lived in a digital world and spend countless hours online—they know no other existence. Nonetheless, these students' engagement in this technological world is frequently more receptive than productive in nature (Foehr 2006). Moreover, in the literature, it has been documented that college students prefer to read and study digitally, as opposed to print, and feel that their performance is better when reading online than offline (Ackerman and Goldsmith 2011; Singer and Alexander 2017). What is particularly interesting about these self-reported judgments is that, in reality, these undergraduates' comprehension performance is significantly better when they read print, except they were being asked quite global or gist questions (e.g., What was the main idea of what you read?). These outcomes suggest that undergraduates may do fine when assessments tap into general understanding but that their performance after working online may suffer when they are required to delve deeply or to be more critical-analytic (Alexander 2014). Such events of poor calibration among students (i.e., inaccurate judgments of performance) have been attributed to the speed-accuracy trade-off (Wickelgren 1977). In effect, because students process faster when online than in print, they assume that they have performed better, without recognizing that the increased speed can contribute to decreased accuracy.

Here again, I do not want to place the burden of information management solely on college students' shoulders. Educators and the educational system are complicit in this situation. For one, especially within the United States, schools have become institutions for test preparation rather than institutions of learning. The specter of high-stakes testing not only casts a long shadow over students, teachers, and school administrators, but it drives curricula where hours upon hours are devoted to preparing students for the upcoming state and national assessments (Au 2007; Nelson 2013). Thus, when it comes to assessment, particularly in grades K-12, the cart is driving the horse. It should come as no surprise, therefore, that college students retain the test preparation mentality when they move into the college classroom.

Yet, there is another pedagogical pattern that must be introduced in this consideration of potential contributors—mentioning versus teaching. Nearly 40 years ago,

Durkin (1978–1979) coined the term “mentioning” to capture the rather superficial level of instruction she observed in elementary classrooms when it came to reading comprehension. Durkin argued that this superficial instruction should not be mistaken for actual teaching, which required focused, systematic pedagogical delivery. I see Durkin’s distinction between mentioning and teaching as even more relevant in contemporary contexts for several reasons. Principally, the amount of information to be covered within college classes has expanded over the ensuing decades. Textbooks written to survey courses can devote little time or space to any one concept or procedure, and instructors tasked with delivering these courses find themselves struggling to “cover” the content. With more and more to teach, the tendency to *mention* is unquestionably strong. As a consequence of these conditions, college students are afforded little time to play with ideas or to reflect deeply on them. Concomitantly, college instructors are driven to sample from the myriad of ideas mentioned during lectures or within the pages of course materials. Thus, the opportunities for students to build knowledge and for instructors to craft assessments that focus on the processes associated with knowledge building are severely hampered.

3.4 What Can Be Done?

It would be unfair for me to lay bare the problems that I see regarding information management and knowledge building in twenty-first-century students without offering any potential resolutions. Therefore, I want to focus here on steps that those within higher education could take in their teaching and in their assessment practices to foster more effective management of information and to forge more opportunities for knowledge building among the students they seek to serve. Although I present these responses in rather simple and straightforward language, I acknowledge that their execution is anything but simple or straightforward. We have not come to this situation overnight. Moreover, the forces at work here are systemic and societal in character. Nonetheless, there are steps that can be taken and should be taken in teaching and assessment at the tertiary level if the goal is to promote the development of “good learners” and not just “good students.”

3.4.1 *Teach More About Less*

If college educators are going to create the educational places where knowledge building can be fostered, then the prevailing tendency to cover massive amounts of content must be set aside. This is what Schmidt et al. (2011) refer to as the *mile-wide and inch-deep* curricular phenomenon. The counterpoint to this mile-wide and inch-deep approach is to teach more about less (Alexander and Knight 1993). That is because effective information management and certainly knowledge building require sufficient time and adequate instructional space. Further, when too much is

crammed into the college course, “mentioning” (Durkin 1978–1979) becomes a necessity. Yet mentioning seems antithetical to the provision of adequate time or space that effective information management and knowledge building demand. In contrast, when more instructional time is devoted to those concepts and procedures that are regarded as central to the domain or discipline being studied, students have greater opportunity to grasp the nature and importance of that content. When this instructional time encompasses occasions for students to apply what they are learning in meaningful ways and to engage in critical discussions and problem solving with peers around that content, then knowledge building may arise more naturally.

This call to teach more about less places the onus on college teachers to determine what within the curriculum is truly core to the field and, thus, deserving of instructional attention. In addition, because of their foundational nature, there is an expectation that core concepts and procedures will be repeatedly encountered within a program of study, allowing students to deepen and broaden their domain or disciplinary knowledge. Likewise, the “less is more” principle should be evident in assessment practices in college courses. As a case in point, it may be possible to administer a multiple-choice test containing 100 questions in a college class and to ascertain what specific facts these students have memorized. In contrast, it is conceivable that much more could be learned about students’ grasp of the target content by posing with a few questions that require students to integrate and synthesize what was taught and to use that knowledge in response to a novel and thought-provoking issue or conundrum. This synergy between instruction and assessment should not only pertain to the content that is core to both but also to the cognitive processes that are valued within the domain or discipline, as will be discussed subsequently.

3.4.2 Devise Assessments That Require Reasoning and Justification

As suggested, changing the character of instruction within college classrooms without altering the character of the assessment would likely work against the goal of effective information management and knowledge building (Pellegrino 2006). However, when assessments work in conjunction with the instructional aims of deepening students’ understanding and arming them with the habits of mind and habits of action that are indicative of knowledge building, learners benefit (Pellegrino et al. 2001). Just as the technologies that seem ubiquitous in postindustrial societies are neither the villain nor the heroine of twenty-first-century education, assessment is not strictly the cause nor the cure for contemporary learning woes. What matters are the weight they are given, their characteristics, and the meaningful information they provide. In effect, if the tests remain at the surface of understanding or only survey content and do not demand that students apply understanding in new ways, justify their interpretations, or critique ideas explored, then information management (even at a rather ineffective level) would suffice.

What I am stating here about the viable role of assessment is nothing new. Others have argued for a strong alignment between the instructional goals and the conceptual and procedural aims of assessment (Gibbs and Simpson 2004; Pellegrino et al. 2001). What is novel about this discussion is that the attributes of effective assessment practices are reframed in terms of knowledge building and information management. Moreover, what this reframing suggests is that the attributes of quality assessments within higher education may merit reexamination in terms of the desired manifestations of knowledge building. For instance, assessments that ask students to apply their understanding in new and novel ways or that require evidence of critical-analytic thinking or justification of responses would entail processes associated with knowledge building.

3.4.3 Explicitly Teach Strategies for Effective Information Management

The instructional mission of institutions of higher learning is not just to teach facts or concepts, but also to facilitate the development of strategic processes that underlie effective information management. There is extensive evidence that students benefit from learning *how* to think, to reason, and to regulate their cognition (e.g., Bransford et al. 1999; Zimmerman 2002). If students must confront the inevitable challenges of navigating the informational waters surrounding them, then they must be equipped with the strategic tools required for such successful navigation. Returning to the proffered description of the processes underlying information management, college students need to be able to (a) ascertain the nature of a given task, (b) focus attention and perception on critical task components, (c) distinguish relevant from irrelevant content within the informational stream, (d) devise strategies for retaining relevant content over time, and (e) effectively apply the information to the well-analyzed task before them. As researchers have garnered from generations of strategy research, it is best to target these underlying processes within the flow of domain-specific instruction and not in any disembodied or overly general manner (Alexander et al. *in press*; Graham and Perin 2007). For example, notions such as *relevant* and *irrelevant* have little value when discussed generically, since the task and context are what give those terms definition. However, when a particular topic is being addressed (e.g., theories of learning, acids versus bases, or biological taxonomic system), there is substantive knowledge that can be brought to bear on the question of what is relevant and what is not (e.g., Oser et al., Chap. 7, in this volume). Further, instructors can guide students to recognize the value of discerning relevance and to arm them with procedures for making such a determination. As this one case suggests, there are ample opportunities for college instructors to promote more effective information management among their students, even as they convey essential course content.

3.4.4 Focus on the Habits of Mind Associated with Knowledge Building

As I have argued throughout this treatise, information management alone is not sufficient to permit twenty-first-century students to become twenty-first-century learners. They must harness the power of knowledge building. If they ever hope to gain competence in any complex field of study, students must at some point do more than deal with the tasks given them by others or be content to allow others to set the standards for evaluation (Alexander 1997; Nandagopal and Ericsson 2012). Specifically, as I stated, knowledge building requires individuals to (a) establish their own intentions for engaging information, (b) identify the importance or salience of that engaged information, (c) gather sufficient evidence that would allow for epistemic judgments about the information encountered, (d) engage in critical or reflective analysis of what is seen or heard, and (e) use that evidence and reflection to determine the veracity or credibility of domain- or task-relevant information.

These processes demand more of learners and the learning environment than is typical within higher education. For one, it is not enough for students to be aware of their personal goals or intentions. There must be space within the curriculum for students to pursue those goals and intentions in relation to the course content (Kulikowich and Alexander 2010). For another, the standards for gathering and judging disciplinary or domain evidence may not be transparent to those still acclimating to a field of study (Bråten et al. 2011; Mason et al. 2011). Therefore, it falls to instructors to enlighten students about pertinent sources of evidence for a field, as well as the criteria to consider in judging those sources (Alexander and the Disciplined Reading and Learning Research Laboratory 2012).

In addition, as the growing body of research on critical-analytic thinking strongly indicates, students cannot be assumed to possess the strategies that result in defensible evidence-based decisions (Alexander 2014; Murphy et al. 2014). As with information management processes, these indicators of knowledge building must be practiced and reinforced within the flow of instruction and within the context of classroom-based assessments. Also, for assessments to be of particular value in knowledge building, they should provide students with useable feedback on performance that may deepen their understanding of central concepts and procedures and resulting judgments as to the veracity or “truthfulness” of what is read or heard (Hattie 1993, 2013; Shavelson et al., Chap. 10, in this volume).

3.5 Concluding Thoughts

I trust that the “discovery” of the serious discrepancy between twentieth-century perceptions of effective teaching and the realities under which twenty-first-century students live and learn has been laid bare. There is no question that there are viable reasons why today’s college students often operate from the mind-set of “just

managing”—managing to figure out what is required of them, managing to fulfill academic expectations without going any deeper in thought or action than necessary, and managing to purge themselves of unnecessary content once immediate needs are fulfilled. Yet, there are tremendous gains to be realized when those habits of mind are reoriented toward knowledge building as well as effective information management—where information is weighed and measured against domain and disciplinary standards in order to test its veracity or credibility—and when effort is exerted toward not simply memorizing content for the short term but to integrate resulting understandings into one’s knowledge base to be used in future and varied contexts (e.g., Oser et al., Chap. 7, in this volume). The process of recasting higher education teaching and assessment in such a way as to promote both effective information management and knowledge building will be demanding and, at times, exhausting for all those involved. But the alternative would be to leave students to flounder in the ocean of information that marks their lives, in and out of classrooms, and to allow the habits of mind indicative of knowledge building to continue to atrophy. Neither of these circumstances is desirable. Thus, those committed to improved teaching and assessment in higher education must act as role models in the face of this challenge and offer their students both structure and opportunity to build knowledge and must do so before these undesirable conditions are allowed to exacerbate.

References

- Ackerman, R., & Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus paper. *Journal of Experimental Psychology: Applied*, 17(1), 18–32.
- 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: JAI Press.
- Alexander, P. A. (2014). Thinking critically and analytically about critical-analytic thinking: An introduction. *Educational Psychology Review*, 26(4), 469–476.
- Alexander, P. A. (2015a). A+ students/c- learners: Education’s report card. *Psychology Today*, American Psychology Association Blog Series. <https://www.psychologytoday.com/blog/psyched/201502/studentsc-learners-education-s-report-card>
- Alexander, P. A. (2015b). *Information management versus knowledge building: Implications for learning and assessment in higher education*. Affiliated meeting of KOKOHs, competence modeling and competence assessment in higher education, Johannes Gutenberg-University Mainz Germany and the American Educational Research Conference, Chicago.
- Alexander, P. A., & Knight, S. L. (1993). Dimensions of the interplay between learning and teaching. *The Educational Forum*, 57, 232–245.
- Alexander, P. A., & Murphy, P. K. (1998). Profiling the differences in students’ knowledge, interest, and strategic processing. *Journal of Educational Psychology*, 90, 435–447.
- Alexander, P. A., & The Disciplined Reading and Learning Research Laboratory. (2012). Reading into the future: Competence for the 21st century. *Educational Psychologist*, 47(4), 1–22. <https://doi.org/10.1080/00461520.2012.722511>.
- Alexander, P. A., Jetton, T. L., & Kulikowich, J. M. (1995). Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology*, 87, 559–575.

- Alexander, P. A., Dumas, D., Grossnickle, E. M., List, A., & Firetto, C. M. (2015). Measuring relational reasoning. *Journal of Experimental Education*, *84*, 119. <https://doi.org/10.1080/00220973.2014.963216>.
- Alexander, P. A., Singer, L., Jablansky, S., & Hattan, C. (2016). Relational reasoning in word and in figure. *Journal of Educational Psychology*, *108*, 1140–1152.
- Alexander, P. A., Grossnickle, E. M., Dumas, D., & Hattan, C. (in press). A retrospective and prospective examination of cognitive strategies and academic development: Where have we come in twenty-five years? In A. O'Donnell (Ed.), *Handbook of educational psychology*. Oxford: Oxford University Press.
- Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267.
- Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the observed learning outcome)*. New York: Academic Press.
- Bok, D. (2009). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton: Princeton University Press.
- Braasch, J. L., Rouet, J. F., Vibert, N., & Britt, M. A. (2012). Readers' use of source information in text comprehension. *Memory and Cognition*, *40*(3), 450–465.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.
- Bråten, I., Strømshø, H. I., & Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learning and Instruction*, *21*(2), 180–192.
- Dinsmore, D. L., & Alexander, P. A. (2012). A critical discussion of deep and surface processing: What it means, how it is measured, the role of context, and model specification. *Educational Psychology Review*, *24*(4), 499–567.
- Dumas, D., Alexander, P. A., Baker, L. M., Jablansky, S., & Dunbar, K. M. (2014). Relational reasoning in medical education: Patterns in discourse and diagnosis. *Journal of Educational Psychology*, *106*(4), 1021–1035.
- Durkin, D. (1978–1979). What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly*, *14*, 481–533.
- Entwistle, N., & Entwistle, D. (2003). Preparing for examinations: The interplay of memorizing and understanding, and the development of knowledge objects. *Higher Education Research & Development*, *22*(1), 19–41.
- Entwistle, N. J., & Peterson, E. R. (2004). Conceptions of learning and knowledge in higher education: Relationships with study behaviour and influences of learning environments. *International Journal of Educational Research*, *41*(6), 407–428.
- Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.
- Foehr, U. G. (2006). *Media multitasking among American youth: Prevalence, predictors and pairings*. Washington, DC: Henry J. Kaiser Family Foundation.
- Franco, G. M., Muis, K. R., Kendeou, P., Ranellucci, J., Sampasivam, L., & Wang, X. (2012). Examining the influences of epistemic beliefs and knowledge representations on cognitive processing and conceptual change when learning physics. *Learning and Instruction*, *22*(1), 62–77.
- Gibbs, G., & Simpson, C. (2004). Does your assessment support your students' learning? *Journal of Teaching and Learning in Higher Education*, *1*(1), 1–30.
- Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476.
- Hattie, J. A. (1993). Measuring the effects of schooling. *SET*, *2*, 1–4.
- Hattie, J. (2013). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

- Hembrooke, H., & Gay, G. (2003). The laptop and the lecture: The effects of multitasking in learning environments. *Journal of Computing in Higher Education*, *15*(1), 46–64.
- Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemporary Educational Psychology*, *25*(4), 378–405.
- Kulikowich, J. M., & Alexander, P. A. (2010). Intentionality to learn, in an academic domain. *Early Education and Development*, *21*(5), 724–743.
- List, A., & Alexander, P. A. (2017). Cognitive affective engagement model of multiple source use. *Educational Psychologist*, *52*, 182. <https://doi.org/10.1080/00461520.2017.1329014>.
- List, A., Grossnickle, E. M., & Alexander, P. A. (2015). Undergraduate students' justifications for source selection in a digital academic context. *Journal of Educational Computing Research*, *54*, 22. <https://doi.org/10.1177/0735633115606659>.
- Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British Journal of Educational Psychology*, *46*(1), 4–11.
- Mason, L., Ariasi, N., & Boldrin, A. (2011). Epistemic beliefs in action: Spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. *Learning and Instruction*, *21*(1), 137–151.
- Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientation and cognitive engagement in classroom activities. *Journal of Educational Psychology*, *80*, 514–523.
- Murphy, P. K., Rowe, M. L., Ramani, G., & Silverman, R. (2014). Promoting critical-analytic thinking in children and adolescents at home and in school. *Educational Psychology Review*, *26*(4), 561–578.
- Nandagopal, K., & Ericsson, K. A. (2012). Enhancing students' performance in traditional education: Implications from expert performance approach and deliberate practice. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *Educational psychology handbook* (Vol. 1, pp. 257–293). Washington, DC: American Psychological Association.
- Nelson, H. (2013). *Testing more, teaching less: What America's obsession with student testing costs in money and lost instructional time*. Washington, DC: American Federation of Teachers.
- Ophir, E., Nass, C., & Wagner, A. D. (2009). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, *106*(37), 15583–15587.
- Pellegrino, J. W. (2006). *Rethinking and redesigning curriculum, instruction and assessment: What contemporary research and theory suggests*. Washington, DC: National Center on Education and the Economy for the New Commission on the Skills of the American Workforce.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.
- Prensky, M. (2001). Digital natives, digital immigrants: Part 1. *On the Horizon*, *9*(5), 1–6.
- Richtel, M. (2010, November 10). Growing up digital, wired for distraction. *The New York Times*, A1.
- Rideout, V. J., Foehr, U. G., & Roberts, D. F. (2013). *Generation M2: Media in the lives of 8-to 18-year-olds*. Washington, DC: Kaiser Family Foundation Study.
- Rosen, L. D., Whaling, K., Rab, S., Carrier, L. M., & Cheever, N. A. (2013). Is Facebook creating “iDisorders”? The link between clinical symptoms of psychiatric disorders and technology use, attitudes and anxiety. *Computers in Human Behavior*, *29*(3), 1243–1254.
- Schmidt, W. H., Houang, R., & Cogan, L. S. (2011). Preparing future math teachers. *Science*, *332*(603), 1266–1267.
- Selwyn, N. (2003). Apart from technology: Understanding people's non-use of information and communication technologies in everyday life. *Technology in Society*, *25*(1), 99–116.
- Singer, L. M., & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education*, *85*(1), 155–172.
- Stanovich, K. E., West, R. F., & Toplak, M. E. (2011). The complexity of developmental predictions from dual process models. *Developmental Review*, *31*(2), 103–118.

- Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41(1), 67–85.
- Wood, E., Zivcakova, L., Gentile, P., Archer, K., De Pasquale, D., & Nosko, A. (2012). Examining the impact of off-task multi-tasking with technology on real-time classroom learning. *Computers & Education*, 58(1), 365–374.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70.
- Zusho, A., Pintrich, P. R., & Coppola, B. (2003). Skill and will: The role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education*, 25(9), 1081–1094.

Part II
Domain-Specific Student Learning
Outcomes – National and International
Perspectives

Chapter 4

Challenges for Evaluation in Higher Education: Entrance Examinations and Beyond: The Sample Case of Medical Education



Christiane Spiel and Barbara Schober

Abstract The present chapter discusses evaluation in higher education using the example of medical education. Concretely, the chapter has three aims: (1) we illustrate potential limitations of “classic” entrance examinations for medical education in addition to their limited scope in terms of academic competencies, using a very detailed evaluation of the Austrian medical school entrance examination as a sample case; (2) we briefly summarize the changes the concepts and goals of entrance examinations in medical education have undergone in recent years. There has been a shift to also include interpersonal and intrapersonal competencies in the medical school admission process, as these competencies are considered decisively relevant for physicians as members of the future health care workforce; and (3) we propose a comprehensive evaluation model for competence-based teaching that begins by defining the competencies students should acquire through higher education. In a second step, the curriculum has to be developed in a way that addresses all of the goal competencies that have been defined. The final step should then be to define the competencies incoming freshmen should possess in order to be successful in university training and specifically in medical education.

4.1 Introduction

The main goal of entrance examinations in higher education is to reliably identify whether candidates are likely to be successful in university training (e.g., Spiel et al. 2007). Aptitude tests in conjunction with grade point average (GPA) were and still are widely used as the method of choice for university admission, and selection is mostly based on knowledge exams (Abbiati et al. 2016). This was also the case in

C. Spiel (✉) · B. Schober

Faculty of Psychology, University of Vienna, Vienna, Austria

e-mail: christiane.spiel@univie.ac.at; barbara.schober@univie.ac.at

medical education for many years, with exams mainly focusing on basic and natural science (Mahon et al. 2013). In recent years, there has been a shift to also include interpersonal and intrapersonal competencies in the medical school admission process, as these competencies are considered decisively relevant for physicians as members of the future health care workforce (Abbiatti et al. 2016; Mahon et al. 2013; Sade et al. 1985). This shift in the conception of entrance examinations can also be attributed to the general shift from teacher-centered education to learner-centered education within the “Bologna Process” (European Commission 2014). However, revising admission practices by adding new sections is a very limited approach. Instead, a comprehensive approach is needed, one that begins by defining the competencies students should acquire through higher education rather than jumping immediately to admission procedures (see also Coates in this volume). In a second step, the curriculum has to be developed in a way that addresses all of the goal competencies that have been defined (Bergsmann et al. 2015). The final step should then be to define the competencies incoming freshmen should possess in order to be successful in university training and specifically in medical education. Such a comprehensive view of competence-based teaching takes serious account of learner-centered education, which is still widely neglected elsewhere (Bergsmann et al. 2015). Consequently, competence-based teaching defined and realized in this way needs a corresponding comprehensive evaluation model that goes far beyond entrance examinations for applicants.

The present chapter discusses evaluation in higher education using the example of medical education. We begin with entrance examinations and move toward a comprehensive evaluation model for competence-based teaching, which we present at the chapter’s conclusion. Concretely, the chapter has three aims: (1) we illustrate potential limitations of “classic” entrance examinations for medical education in addition to their limited scope in terms of academic competencies, using a very detailed evaluation of the Austrian medical school entrance examination as a sample case; (2) we briefly summarize the changes the concepts and goals of entrance examinations in medical education have undergone in recent years; and (3) we propose a comprehensive evaluation model for competence-based teaching.

4.2 Evaluation of the Entrance Examination for Medical Education in Austria

In Austria, by law, access to public universities had always been unrestricted, with the exception of arts universities (Perthold-Stoitzner 2016). The only requirement was the school-leaving examination (similar to the British A-level). In 2006, a decision of the European Court required Austria to treat all applicants for university studies from the European Union (EU) equally (Spiel et al. 2008). As a consequence, a high number of German applicants for university study programs in Austria were expected. To cope with this situation, Austrian universities were allowed to establish entrance examinations in all university subjects in which there was a *numerus clausus* [enrolment limit] in Germany.

As medical education was one of the most popular degree programs in both Austria and in Germany, the three medical universities in Austria established entrance examinations beginning in 2006 (under high time pressure). The results of these entrance examinations revealed dramatic differences between female and male applicants. In 2007, 57.3% of applicants ($n = 3623$) were females, but only 44.0% of the candidates accepted on the basis of their test scores were females. This was true despite the fact that previous studies of medical education have clearly shown that sex is not a valid predictor of success (Mitterauer et al. 2007). While females are less successful than males in the early phases of medical education, in the long run, sex differences in performance diminish, and more females receive medical degrees than males. In order to explain these differences, the Austrian Federal Ministry for Science and Research commissioned an evaluation of the entrance examination (Spiel et al. 2008). The focus of this evaluation went beyond the assessment of classical psychometric properties and placed particular emphasis on the examinations' gender fairness.

Two of the medical universities in Austria used the Swiss aptitude test for medical education (Eignungstest für Medizinische Studiengänge, EMS) as their selection method, while a new test measuring knowledge of natural sciences was developed at the third university. As the results were very similar, we only present results for the EMS in the following paper. The EMS is based on the German selection test for medical education (Test für Medizinische Studiengänge, TMS) developed in the 1970s. According to its authors (Trost et al. 1998), the TMS was developed to assess candidates' abilities with regard to the first, basic phase of medical training where the focus is on natural sciences. The TMS has been shown to have good predictive validity (correlation with final grades at the end of the first, basic phase of medical education $r = 0.45$, in conjunction with GPA $r = 0.57$; Trost et al. 1998). The EMS is a slightly modified version of the TMS that has been used in Switzerland since 1998 (e.g., Hänsgen and Spicher 2007). It consists of 10 subtests with a total of 198 multiple-choice items (five answer alternatives, one correct solution; consequently a 20% guessing probability). The main focus is on deductive reasoning in medicine and the natural sciences. According to its description, the EMS measures competencies developed over a long time, which cannot be trained in a short period.

Entrance examinations should fulfill several criteria. There is wide agreement that objectivity, reliability, validity, and efficiency are the most important psychometric properties entrance examinations should have (Spiel et al. 2007), with predictive validity as the most relevant criterion. Furthermore, it should not be possible to fake them. Additionally, several psychological and educational criteria have been proposed for entrance examinations (Spiel et al. 2007). Besides such criteria as transparency and acceptance, the definition of the qualification profile and the fairness of the examination are seen as very important here. The qualification profile should cover all competencies necessary for success in university training and – particularly in the case of medicine – also for success in the workforce (to the extent that these competencies are not trained within the curriculum). All of these competences should be assessed in the entrance examination. An entrance examination is considered fair when there is no systematic and unjustified discrimination against specific groups of applicants, for example, on the basis of sex (Spiel et al. 2007). “Unjustified” here means that poor results are caused by reasons other than lower eligibility.

The evaluation was conducted using data from all applicants for medical education at the two universities using the EMS in 2007. The sample consists of 3623 applicants (2075 females); 1936 were Austrians, and 1366 were Germans. The mean age was 20.7 years (standard deviation = 3.2 years). The applicants' scores on the EMS and data from a short questionnaire about their family and school background formed the basis of the evaluation.

Analyses of the psychometric properties of the EMS showed relatively low reliability scores for the subtests (between 0.56 and 0.75) in relation to the number of items (one test had 18 items, while all others had 20 items), with the items having very low discriminatory power (all below 0.3). Objectivity was obviously high as the EMS is a paper-and-pencil multiple-choice test (Spiel et al. 2008). Since the evaluation took the form of a cross-sectional study, predictive validity could not be addressed. Applicants were given 5 h 15 min to complete the EMS. Due to the high number of applicants, conference halls had to be rented to conduct the assessment, and there were a large number of people responsible for dealing with the applicants.

In evaluating the fairness of the EMS in terms of sex, we addressed the so-called accounting fairness. Concretely, we analyzed whether the items in each subtest measured the same competence or knowledge in both males and females using a Rasch model (e.g., Kubinger 1989; Fischer 1995). The reason for applying this procedure was that test scores were used to describe the applicants' competencies. If and only if the Rasch model holds can test scores be used to describe participants' competencies. We applied the Andersen chi-squared test (Kubinger 1989) in the analyses. Results showed that in only three out of the ten subtests did all items measure the same competence in males and females. In one subtest, more than 50% of the items measured different things in males and females. For the other six subtests, an average of 21% of the items was shown to not measure the same competencies in males and females.

In addition, we analyzed the relation between EMS score and GPA in mathematics and the natural sciences while taking sex into account. A clear relation between average grades in mathematics and the natural sciences and test scores could be observed for both sexes, such that better grades were associated with higher EMS scores. However, we also observed dramatic sex differences in the relationship between test scores and grades¹ (see Fig. 4.1). Females had significantly lower test scores than males with equivalent school grades.

There are a number of studies showing that the higher grades achieved by girls do not correspond to their performance in objective achievement tests. For example, Kenney-Benson and colleagues observed that although girls are awarded higher grades in mathematics than boys, their achievement test scores in mathematics are not higher. Moreover, the grade advantage held by girls over boys intensifies over time (Kenney-Benson et al. 2006). Similar results were found in a Swedish study (Nycander 2006). These astonishing results can be explained by the large body of literature on gender stereotypes and the consequences these stereotypes have for

¹In Austria school grades vary from 1 (excellent passed) to 5 (failed).

teachers' behavior and expectations concerning boys' and girls' competencies in mathematics and the natural sciences as well as their work ethics (e.g., Jones and Myhill 2004; for an overview, see Kollmayer et al. 2016). This means, the evaluation identified fairness issues to the detriment of females not only in the EMS, with its limited focus on basic and natural science, but also in earlier school education.

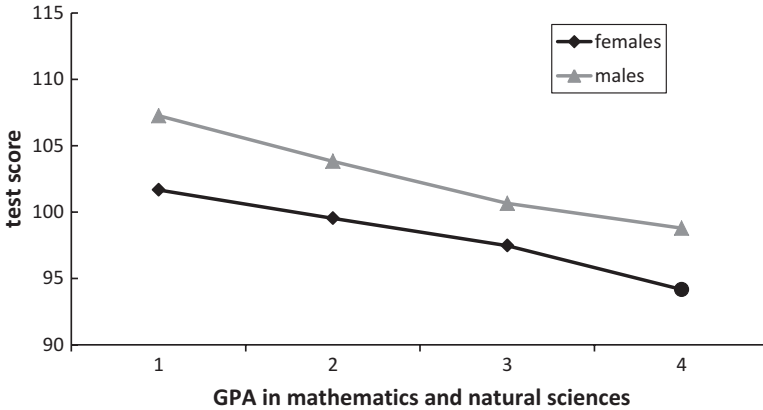


Fig. 4.1 Relation between grades and test scores for females and males

Obviously, an entrance exam cannot solve the problems of schools, but if the competence profile measured in the exam were to cover all of the competence areas needed in the clinical phase of medical school and later in the workforce, including interpersonal and intrapersonal competencies, this problem could be reduced. Females can be expected to outperform males in such competencies – again as a result of gender stereotypes (Kite et al. 2008). The improvements to medical school entrance examinations described in the following section move in exactly this direction.

4.3 Entrance Examinations in Medical Education: New Developments

Entrance examinations have a long tradition in medical education. Even as early as the 1910 Flexner Report (Mahon et al. 2013), it was argued that future physicians should possess a minimum threshold of knowledge in the basic and natural sciences. Consequently, selection for most medical schools worldwide was and still is based on knowledge exams (Abbiati et al. 2016), with the EMS serving as a prototypical example for this. The Medical College Admission Test (MCAT), for example, has been the tool of choice for more than 80 years not only to measure medical school applicants' mastery of scientific content, in conjunction with their grade point average, but also as a reliable predictor of performance on medical school and initial licensure examinations (Mahon et al. 2013). The usage of such selection

methods is supported by a consistent body of evidence showing that cognitive measures of high school performance such as GPA or scores on cognitive tests predict academic success during higher education in general (e.g., Spiel et al. 2007) and specifically in medical education (e.g., Sawyer 2013).

The potential problems and shortcomings of focusing solely on knowledge-based exams were widely neglected for many years. Nevertheless, it is not just academic competencies but also interpersonal and intrapersonal competencies that are necessary for university graduates in general and for future physicians in particular (Abbiati et al. 2016; Mahon et al. 2013; Patterson et al. 2016; Spiel et al. 2008). Obviously, measuring these competencies reliably is much more difficult than measuring cognitive performance. A further shortcoming is that research on selection methods has been mostly cross-sectional and has tended to focus on reliability estimates rather than validity (Patterson et al. 2016). When longitudinal approaches are applied and predictive validity is analyzed, most studies use grades (often those in the early years of the study) as the criterion measure. However, this type of criterion measure is very similar to the tasks required in entrance examinations, and high scores for predictive validity might therefore be artificial (Spiel et al. 2008). There is a lack of studies linking entrance examinations to measures of clinicians' competencies (Patterson et al. 2016). Last but not least, there are methodological problems in estimating the predictive validity of entrance examinations, as the values on the criterion variable are available only for selected applicants (Pfaffel et al. 2016). This range restriction problem is widely neglected in selecting future physicians but also more broadly.

In recent years, a consistent body of literature has highlighted the importance of a broader approach to selecting candidates for higher education, particularly candidates for medical school (Abbiati et al. 2016; Mahon et al. 2013; see also Sade et al. 1985; Spiel et al. 2007, 2008). The selection of future physicians is increasingly seen as a key component of health care reform, and consequently it is becoming more and more important to select students for medical school admission who will be superior physicians, rather than those who will be excellent medical students in terms of high performance on thematically narrow tests in the university context (Mahon et al. 2013; see also Sade et al. 1985). Academic achievement alone has been shown to poorly correlate with physicians' clinical performance (Sade et al. 1985). A systematic review by Patterson et al. (2016) showed that performance on different selection methods may differentially predict performance at different stages of medical education and clinical practice. As a consequence, the authors argue first for more longitudinal studies focusing on predictive validity that follow students throughout the course of their careers in education, training, and practice. Furthermore, they propose that there should be an increased focus on and value attributed to nonacademic attributes and skills in medical school selection, such as the capability to lead multidisciplinary teams and building a culture of "everyday" innovation in an environment of reduced resources. Additionally, the authors recommend a focus on attracting a wider selection pool and recruiting a more diverse workforce. Mahon et al. (2013) draw similar conclusions.

The shift in entrance examinations to a broader profile of measured competencies corresponds with the shift from teacher-centered education to more learner-centered education (Reynolds and Miller 2013) and to competence-centered curricula (Wesselink et al. 2010). The demand for such a reform was especially evident in medical education (Cantor et al. 1991). As a result of the “Bologna Process” (European Commission 2014), competence-based teaching in higher education has become a significant goal. The German program Modeling and Measuring Competencies in Higher Education (KoKoHs) has made particularly significant contributions to defining competencies and developing corresponding theoretical and measurement models (Blömeke et al. 2013; Zlatkin-Troitschanskaia et al. 2016).

According to Mahon et al. (2013), the shift to competence-based medical education is leading to a parallel shift toward competency-based admission. In 2013, the Association of American Medical Colleges (AAMC) identified the most desirable interpersonal and intrapersonal competencies that medical schools expect of incoming students (Mahon et al. 2013). Furthermore, the AAMC redesigned the MCAT exam (beginning in 2015). Furthermore, the AAMC worked together with other associations to define four interprofessional competencies that students of health professions should acquire over the course of their training: values and ethics, an understanding of roles and responsibilities, interprofessional communication, and teamwork (Interprofessional Education Collaborative Expert Panel 2016).

Obviously, these developments represent significant changes to the goals of medical education in general and the admission process in particular. Nevertheless, substantial improvement is only possible when the competencies that students are expected to acquire within the curriculum are systematically combined with the competencies incoming students should possess as well as competence-based teaching as the bridge between them in a comprehensive approach. To the best of our knowledge, this has not yet been done. Thus, we present such a comprehensive approach in the third section of this chapter.

4.4 An Evaluation Model for Competence-Based Teaching in Medical Education and Beyond

The orientation toward competence-based teaching in higher education and in medical education in particular requires new evaluation concepts that overcome the limitations of existing approaches. Bergsmann et al. (2015) describe three limitations: (1) existing evaluation instruments (e.g., course evaluation) mostly focus on single student competencies, but competence-based education requires concepts and methods for the evaluation of all competencies students should acquire within a concrete curriculum, for example, that of medical education; (2) existing evaluation approaches mostly focus on specific aspects of the teaching process (e.g., single courses, the context), while competence-based higher education requires a comprehensive view of competence-based teaching and a corresponding form of systematic evaluation; (3) most evaluations of teaching in higher education focus on status

assessments without considering the needs of the stakeholders of higher education institutions. Therefore, Bergsmann et al. (2015) recommend a participatory evaluation approach (Cousins and Chouinard 2012; Hansen et al. 2013) that includes relevant stakeholders in the evaluation process. A similar recommendation was made by Leonard and colleagues (Leonard et al. 2016). In this section, we present a comprehensive evaluation model for competence-based teaching that overcomes these limitations. As a prerequisite, we briefly explain our understanding of a competence-based teaching model (Bergsmann et al. 2015).

In educational contexts, the theoretical concept of competence has been approached from different angles (Klieme 2004; see also Klieme et al. 2008; Weinert 1999). One stems from the field of linguistic development and socialization (e.g., Chomsky 1986; Habermas 1981), another from education (e.g., Roth 1971), and a third from psychology (McClelland 1973). Therefore, a large body of studies have presented and discussed different definitions, models, and measurement approaches. Within higher education, the KoKoHs research program financed by the German Federal Ministry for Education and Research has been very influential and innovative in terms of definitions, models, and measurement in competence research (e.g., Blömeke et al. 2013; Zlatkin-Troitschanskaia et al. 2017; for a definition of competence, see e.g., Blömeke et al. 2015). In this chapter, we focus on competencies from an evaluation perspective. Consequently, our proposed competence-based teaching model needs to contain the three dimensions described by Klieme and Leutner and their research group (Hartig et al. 2008; Koeppen et al. 2008): (a) competence areas or a competence structure (such as personal competencies, professional competencies, scientific competencies, etc.); (b) competence levels, which specify the degree of expertise from a basic to a more advanced/professional level; and (c) competence development, which means that competencies can and should be improved upon in higher education in general and particularly in medical education (Blömeke et al. 2013; Hartig et al. 2008). It is expected that students holding a master's degree have a higher competence level than students holding a bachelor's degree and that these bachelor's degree holders are in turn on a higher level than incoming first years.

In developing a competence-based teaching model, the first step is to define the competencies students should acquire through higher education (Bergsmann et al. 2015). These competencies are called "ideal student competencies." In a next step, a curriculum has to be developed that addresses all of the ideal competencies that have been defined. For example, if "communication competencies" are defined as a relevant competence area for medical education, then the curriculum has to contain courses addressing these competencies on the level of both knowledge and skill (Bergsmann et al. 2015). After a curriculum has been established, teaching methods and exam formats can be derived. These should be appropriate for fostering a competence-based learning process as well as learning strategies students should apply (e.g., perspective taking for communication competencies). Such a competence-based teaching process should result in "real student competencies" that are actually acquired by students. Obviously, context variables as student-teacher ratio influence the teaching process. Finally, the competencies incoming

freshmen should possess to be successful in medical training and to be prepared for the workforce have to be defined. Here, it is very important to differentiate between competencies which should be trained and developed within the curriculum and competencies which are necessary prerequisites for successful study, such as competencies in self-regulated learning (OECD 2013a, b). The competence-based teaching model described here is shown in Fig. 4.2.

The corresponding comprehensive evaluation model comprises the following three steps (see Fig. 4.2). The model is slightly modified from Bergsmann et al. (2015) and extended by adding entrance examinations. The three steps are (1) evaluation of the theoretical competence model, (2) evaluation of the teaching process, and (3) evaluation of the entrance examination. The aim of the first step is to evaluate whether the theoretical competence model specifying ideal student competencies meets quality criteria derived from competence research. Consequently, the task for evaluators is to “determine whether the theoretical competence model is well-defined” (Fig. 4.2, Task 1). This step is explained in detail by Bergsmann et al. (2015).

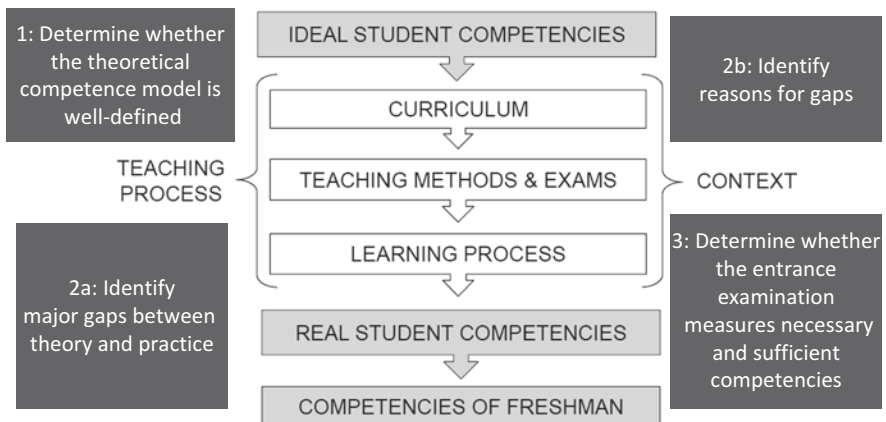


Fig. 4.2 Competence-based teaching model and tasks within a comprehensive evaluation model

The second step for evaluators is to “find out whether there are major gaps between theory and practice” (Fig. 4.2, Task 2a). Concretely, the evaluation here has two aims. The first is to determine whether there is a gap between the ideal level of student competencies and the level of competencies fostered by the teaching process. The second is to determine whether there is a gap between the ideal level and the real level of student competencies. An evaluation of this kind was conducted by Spiel and colleagues (Spiel et al. 2006; see also Spiel et al. 2013; Schober et al. 2004). Their evaluation of medical education at the University of Graz involved four groups of participants: students, university teachers, graduates, and their supervisors. While university teachers and supervisors were asked to assess the ideal level of competencies students should acquire through medical education, students and graduates were asked to assess what was really imparted (thus addressing the first aim of this evaluation step). Additionally, students and graduates assessed the competencies they had acquired through medical education, while university teachers

and supervisors assessed the competencies of their particular learning group, either students or graduates (the second aim of this evaluation step). Results showed considerable gaps between ideal and real competencies in both instances (Spiel et al. 2006). Consequently, the next task for evaluators in this case would be to “identify reasons for gaps” (Fig. 4.2, Task 2b) and conduct a detailed evaluation of the teaching process. Bergsmann et al. (2015) describe how to perform this evaluation task and how to apply a participatory evaluation approach.

The third step of the comprehensive evaluation is the evaluation of the entrance examination. In this step, the evaluators’ task is to “determine whether the entrance examination measures necessary and sufficient competencies” (Fig. 4.2, Task 3). Again, this evaluation has an ideal and a real part. The ideal part is the definition of the profile for incoming students, which should cover all competencies that are necessary for success in medical education and in the workforce. However, it is very important to distinguish between competencies that should be promoted or developed within medical education and competencies that are needed in advance. Only the latter should be assessed in the entrance examination. As described in the first part of the chapter, the entrance examination has to fulfill the requisite psychometric properties and must be fair. To compare the ideal and real elements of the entrance examination, its prospective validity must be assessed (e.g., Spiel et al. 2007; Pfaffel et al. 2016). If the prospective validity score is low, a detailed analysis of potential reasons for this has to be performed, as was the case for the evaluation of the teaching process.

In summary, we strongly recommend that the establishment of a competence-based teaching model be accompanied by a corresponding comprehensive evaluation model, in medical education as well as in other disciplines. The assessment of learning outcomes in higher education can only systematically lead to substantial quality improvement if it is part of an integrative approach: What competencies should students acquire? How can we achieve them through our curriculum? What “equipment” do students need from the beginning on? Only when these questions are answered can the evaluation and corresponding assessments be developed. Entrance exams and their evaluation are only one step in such a comprehensive approach. Evaluation that contributes to improvement goes far beyond this, as we have shown in this chapter. The implementation of comprehensive evaluation models like the one we described here can be useful for diverse stakeholder groups within universities (Bergsmann et al. 2015): (1) rectorate, academic senate, and curriculum commissions can make evidence-based decisions on enhancing curricula and improving teaching quality; (2) university teachers can improve their teaching on the basis of the results of the evaluation; and (3) students can be provided feedback about their own competence profile. The biggest challenge in implementing the evaluation model is building stakeholders’ trust in the evaluation model and in the implementation process as a whole. Therefore, the application of a participatory evaluation approach is strongly recommended.

References

- Abbiati, M., Baroffio, A., & Gerbase, M. W. (2016). Personal profile of medical students selected through a knowledge-based exam only: Are we missing suitable students? *Medical Education Online*, 21, 1–10. <https://doi.org/10.3402/meo.v21.29705>.
- Bergsmann, E., Schultes, M. T., Winter, P., Schober, B., & Spiel, C. (2015). Evaluation of competence-based teaching in higher education: From theory to practice. *Evaluation and Program Planning*, 52, 1–9. <https://doi.org/10.1016/j.evalprogplan.2015.03.001>.
- Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., & Fege, J. (2013). *Modeling and measuring competencies in higher education*. Rotterdam: Sense Publishers.
- Blömeke, S., Gustafsson, J. E., & Shavelson, R. J. (2015). Beyond dichotomies. *Zeitschrift für Psychologie*, 223(1), 3–13. <https://doi.org/10.1027/2151-2604/a000194>.
- Cantor, J. C., Cohen, A. B., Barker, D. C., Shuster, A. L., & Reynolds, R. C. (1991). Medical educators' views on medical education reform. *JAMA*, 265(8), 1002–1006.
- Chomsky, N. (1986). *Language and mind*. New York: Harcourt, Brace & World.
- Cousins, J. B., & Chouinard, J. A. (2012). *Participatory evaluation up close: An integration of research-based knowledge*. Charlotte: Information Age Publishing.
- European Commission. (2014). *The Bologna process and the European higher education area*. Retrieved from http://ec.europa.eu/education/policy/higher-education/bologna-process_en.htm. Accessed 19 Jan 2017.
- Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 15–38). New York: Springer.
- Habermas, J. (1981). *Theorie der kommunikativen Kompetenz* (Vols. 1 & 2). [Theory of communicative competency]. Frankfurt am Main: Suhrkamp.
- Hansen, M., Alkin, M. C., & Wallace, T. L. (2013). Depicting the logic of three evaluation theories. *Evaluation and Program Planning*, 38, 34–43. <https://doi.org/10.1016/j.evalprogplan.2012.03.012>.
- Hänsgen, K.-D., & Spicher, B. (2007). *EMS Eignungstest für das Medizinstudium 2007 – Bericht 13 über die Durchführung und Ergebnisse 2007*. [EMS entrance exam for medical education 2007 – Report 13 on the implementation and results in 2007]. Freiburg: Center for Test Development and Diagnostics.
- Hartig, J., Klieme, E., & Leutner, D. (2008). *Assessment of competencies in educational contexts*. Göttingen: Hogrefe.
- Interprofessional Education Collaborative. (2016). *Core competencies for interprofessional collaborative practice: 2016 update*. Washington, DC: Interprofessional Education Collaborative. Retrieved from https://ipecollaborative.org/uploads/IPEC-2016-Updated-Core-Competencies-Report_final_release_PDF
- Jones, S., & Myhill, D. (2004). 'Troublesome boys' and 'compliant girls': Gender identity and perceptions of achievement and underachievement. *British Journal of Sociology of Education*, 25(5), 547–561.
- Kennedy-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology*, 42(1), 11–26.
- Kite, M. E., Deaux, K., & Haines, E. L. (2008). Gender stereotypes. In *Psychology of women: A handbook of issues and theories* (Vol. 2, pp. 205–236). Westport: Praeger.
- Klieme, E. (2004). Was sind Kompetenzen und wie lassen sie sich messen? [What are competences and how can they be measured?]. *Pädagogik*, 56, 10–13.
- Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 3–22). Göttingen: Hogrefe.

- Koepfen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Journal of Psychology, 216*, 61–73. <https://doi.org/10.1027/0044-3409.216.2.61>.
- Kollmayer, M., Schober, B., & Spiel, C. (2016). Gender stereotypes in education: Development, consequences, and interventions. *European Journal of Developmental Psychology, 1*–17. <https://doi.org/10.1080/17405629.2016.1193483>.
- Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie. [Current state and critical evaluation of probabilistic test theory]. In K. D. Kubinger (Ed.), *Moderne Testtheorie – Ein Abriss samt neuesten Beiträgen* (pp. 19–83). Weinheim: Beltz.
- Leonard, S. N., Fitzgerald, R. N., & Riordan, G. (2016). Using developmental evaluation as a design thinking tool for curriculum innovation in professional higher education. *Higher Education Research & Development, 35*(2), 309–321. <https://doi.org/10.1080/07294360.2015.1087386>.
- Mahon, K. E., Henderson, M. K., & Kirch, D. G. (2013). Selecting tomorrow's physicians: The key to the future health care workforce. *Academic Medicine, 88*(12), 1806–1811.
- McClelland, D. C. (1973). Testing for competence rather than for intelligence. *American Psychologist, 28*, 1–14.
- Mitterauer, L., Frischenschlager, O. & Haidinger, G. (2007). Sex differences in study progress at Medical University of Vienna. *GMS Zeitschrift für Medizinische Ausbildung, 24*(2), Doc111.
- Nycander, M. (2006). *Pojkars och flickors betyg* [Grades received by boys and girls]. Uppsala: Uppsala University, Department of Education.
- OECD. (2013a). *AHELO feasibility study report: Volume 1 – Design and implementation*. Paris: OECD. Retrieved from: www.oecd.org/edu/ahelo
- OECD. (2013b). *AHELO feasibility study report: Volume 2 – Data analysis and national experience*. Paris: OECD. Retrieved from: www.oecd.org/edu/ahelo
- Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., & Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education, 50*(1), 36–60. <https://doi.org/10.1111/medu.12817>.
- Perthold-Stoitzner, B. (2016) *Universitätsgesetz 2002*. [2002 university law]. (4th. Ed.) Vienna: Manz-Verlag.
- Pfaffel, A., Kollmayer, M., Schober, B., & Spiel, C. (2016). A missing data approach to correct for direct and indirect range restrictions with a dichotomous criterion: A simulation study. *PLoS One, 11*(3), e0152330.
- Reynolds, W. M., & Miller, G. E. (2013). Educational psychology: Contemporary perspectives. In I. B. Weiner, W. M. Reynolds, & G. E. Miller (Eds.), *Handbook of psychology, educational psychology* (Vol. 7, pp. 1–22). Hoboken: Wiley.
- Roth, H. (1971). *Pädagogische Anthropologie* (Vol. 2). [Pedagogical anthropology]. Hannover: Schroedel.
- Sade, R. M., Stroud, M. R., Levine, J. H., & Fleming, G. A. (1985). Criteria for selection of future physicians. *Annals of Surgery, 201*(2), 225.
- Sawyer, R. (2013). Beyond correlations: Usefulness of high school GPA and test scores in making college admissions decisions. *Applied Measurement in Education, 26*(2), 89–112.
- Schober, B., Spiel, C., & Reimann, R. (2004). Young physicians' competences from different points of view. *Medical Teacher, 26*(5), 451–457.
- Spiel, C., Schober, B., & Reimann, R. (2006). Evaluation of curricula in higher education: Challenges for evaluators. *Evaluation Review, 30*, 430–450.
- Spiel, C., Litzenberger, M., & Haiden, D. (2007). Bildungswissenschaftliche und psychologische Aspekte von Auswahlverfahren. [Educational science and psychological aspects of selection procedures]. In C. Badelt, W. Wegscheider, & H. Wulz (Eds.), *Hochschulzugang in Österreich* (pp. 479–552). Graz: Grazer Universitätsverlag.
- Spiel, C., Schober, B., & Litzenberger, M. (2008). *Evaluation der Eignungstests für das Medizinstudium in Österreich*. [Evaluation of the entrance examinations for medical education in Austria]. Vienna: University of Vienna, Faculty of Psychology.

- Spiel, C., Schober, B., & Reimann, R. (2013). Modeling and measurement of competencies in higher education: The contribution of scientific evaluation. In O. Zlatkin-Troitschanskaia & S. Blömeke (Eds.), *Modeling and measurement of competencies in higher education* (pp. 195–206). Rotterdam: Sense Publishers.
- Trost, G., Blum, F., Fay, E., Klieme, E., Maichle, U., Meyer, M., et al. (1998). *Evaluation des Tests für medizinische Studiengänge (TMS): Synopse der Ergebnisse* [Report]. [Evaluation of the test for medical education (TMS): Synopsis of the results]. Bonn: Institute for Test und Giftedness Research.
- Weinert, E. F. (1999). *Concepts of competence. Contribution within the OECD project definition and selection of competencies: Theoretical and conceptual foundations (DeSeCo)*. Munich: Max Planck Institute for Psychological Research.
- Wesselink, R., Dekker-Groen, A. M., Biemans, H. J., & Mulder, M. (2010). Using an instrument to analyse competence-based study programmes: Experiences of teachers in Dutch vocational education and training. *Journal of Curriculum Studies*, 42(6), 813–829. <https://doi.org/10.1080/00220271003759249>.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Toepper, M., & Lautenbach, C. (2016). *Messung akademisch vermittelter Kompetenzen von Studierenden und Hochschulabsolventen. Ein Überblick zum nationalen und internationalen Forschungsstand*. Wiesbaden: Springer.
- 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.

Chapter 5

Teachers' Judgments and Decision-Making: Studies Concerning the Transition from Primary to Secondary Education and Their Implications for Teacher Education



Sabine Krolak-Schwerdt, Ineke M. Pit-ten Cate, and Thomas Hörstermann

Abstract Accuracy in assessing academic achievement and potential is a core component of teachers' diagnostic competence. Large-scale studies in the Luxembourgish and German educational systems show that teachers' secondary school track decisions are biased by a student's social background. Therefore, biased assessment of students may contribute to the social inequalities observed in secondary schools in both countries. Within a social cognitive framework of dual-process theories, bias is explained by heuristic information processing, which, in contrast to information-integrating processing, relies on stereotype-based expectations to form judgments about students. A series of experimental studies investigated the information processing strategies of teachers, identifying a low accountability of the decision setting and a high consistency of student information as key moderators that promote stereotype-based information processing strategies in teachers' school track decisions. Results on intervention modules gave insights how to increase diagnostic competence in teacher education programs.

5.1 Introduction

One of teachers' key skills is the ability to adequately judge students' achievements and learning potential. Assessments of student achievement, such as grading, assessment of competence levels, or decisions upon the secondary school type students should attend, are considered key tasks of the teacher's profession. Such assessments not only guide educational pathways of students but may also impact

S. Krolak-Schwerdt (Deceased)

I. M. Pit-ten Cate (✉) · T. Hörstermann
University of Luxembourg, Luxembourg, Luxembourg
e-mail: ineke.pit@uni.lu; thomas.hoerstermann@uni.lu

future occupational opportunities. Therefore, these assessments should fulfill high measurement quality. However, research on the assessment competence of teachers shows that teachers' judgments are neither reliable nor valid (e.g., Ingenkamp and Lissmann 2008), whereby nonacademic variables, such as the gender or social background of the students, influence assessments (Oakes and Guiton 1995; Parks and Kennedy 2007; Pietsch and Stubbe 2007; Weiss 1989). The discussion about the importance of the influence of such nonacademic aspects on teachers' judgments has gained momentum through the international school performance studies Programme for International Student Assessment (PISA) and Progress in International Reading Literacy Study (PIRLS, known in Germany under the term Internationale Grundschul-Lese-Untersuchung, IGLU) (Bamberg et al. 2010; Baumert et al. 2001; Bos et al. 2004; OECD 2010).

Among teachers' judgment tasks with the highest impact are recommendations for school placement or tracking decisions. In many European countries, entry into secondary schools is based on a selection process whereby teachers, together with other professionals and parents, decide which track would be most suitable for the student. Although opinions regarding the validity of school tracking systems vary (see Sect. 5.2.), teachers' ability to accurately assign students to different tracks is not only important for the student's direct educational pathway but may also have a long-lasting effect on students' future careers and the quality of adult life (Dustmann 2004; Kaufman and Rosenbaum 1992; Schalke et al. 2013). School procedures generally reflect the notion that students' academic achievements (i.e., school grades and results of standardized achievement tests) are the best predictors of tracking decisions (Ditton and Krüsken 2009; Haller 1985; Hallinan and Dame 1996; Oakes and Guiton 1995). However, empirical findings show that nonacademic variables also affect decisions, such that students with an immigrant background or lower socioeconomic status (SES) have less chance to receive a recommendation for the highest track (Bertemes et al. 2013; Klapproth et al. 2012; Oakes and Guiton 1995; OECD 2010).

Therefore, the aim of the current research was to obtain deeper insight into teachers' processing of student information and their corresponding decision-making. We chose to investigate teachers' tracking decisions as the type of assessment for several reasons. First, assigning students to different educational pathways and school tracks pertains an important task of the teachers' profession. Second, these tracking decisions strongly contribute to students' future educational and professional careers as already noted. Finally, teachers are often the main decision-makers when it comes to tracking and school placement (Ansalone and Biafora 2004).

This chapter provides an overview of our research on teachers' tracking recommendations and decision-making processes. More specifically the chapter describes findings from consecutive projects aimed to investigate (a) to what extent teachers' judgments and associated decisions may be biased, (b) the extent to which biases are associated with decision (in)accuracy, and (c) the extent to which different interventions can improve decision accuracy by reducing bias.

5.2 Research on the Effects of Tracking

Especially in the context of large-scale comparative studies such as PISA, structural differences between school systems and their potential effect on students' learning outcomes have come into the focus. One structural and repeatedly discussed characteristic of educational systems is tracking. Tracking corresponds to an ability-based grouping of students in different study programs at a specified moment in their school career. Tracking has been criticized by several researchers (for a review, see Gamoran 1992; Van de Werfhorst and Mijs 2010), because it is hypothesized that tracking may lead to phenomena of social and ethnic segregation and may be harmful to the personal and later professional development of students, especially those oriented toward the lower tracks (Alpert and Bechar 2008). Furthermore, it is argued that tracking decisions taken early in the school career constitute a very uncertain prediction of expected learning outcomes (Oakes 2005). On the other hand, support to the tracking systems is given by the hypothesis that the creation of more homogeneous classrooms helps in the organization of ability-adapted learning settings and helps especially high-ability students to progress more rapidly (Kulik and Kulik 1982).

It should be noted that the potential negative effects of tracking can have different origins that may produce their effects in combination. First, there might be negative effects associated with the tracking itself. More specifically, there is a segregation of students based on their proficiency profiles in order to construct more homogeneous learning groups, whereby it is hypothesized that heterogeneous learning groups lead to better overall learning outcomes than homogeneous learning groups. Second, there might be negative effects associated with incorrect tracking decisions. These effects might be more severe when decisions are taken early, at a moment when the empirical basis for the evaluation of the learning potential of a student is still quite uncertain.

Many European educational systems select students for different secondary school tracks according to their achievement level. For example, in the German and Luxembourgish school system, students are grouped into three hierarchical tracks based on their academic achievement. The tracks differ in the type of qualification the students will be able to acquire. Students in the highest school track graduate with a general qualification for university entrance, whereas students in the middle track can acquire different qualifications for professional and vocational education. A student from the lowest school track can graduate with a qualification for vocational education in only limited job areas. Teachers are the main decision-makers in these countries, as either they decide which secondary school type their students will attend or they recommend a school type, which is generally accepted by parents and students.

From the results of large-scale studies on student achievement, it is well known that minority students with immigration backgrounds are underrepresented in the

highest track, whereas they are overrepresented in the lowest track (e.g., Baumert and Schümer 2002, for Germany; Martin et al. 2008, for Luxembourg). However, this phenomenon is not restricted to European countries; in the USA, minority students are also overrepresented in vocational tracks (Ekstrom et al. 1988; Oakes 2005), in lower level tracks (Ansalone 2001; Lucas 1999, 2001), and in classes associated with low-skilled jobs (Oakes and Guiton 1995). Although minority students with immigrant backgrounds often have lower achievement levels, they are disadvantaged even when academic achievement is controlled for (Bamberg et al. 2010; Dauber et al. 1996; Klapproth et al. 2012).

As teachers are the main decision-makers, not only in regard to the tracking decision itself but also concerning the evaluation of school performance and grade assignment, they may contribute to these inequalities. To this extent, empirical findings have demonstrated that assessments are affected by the students' race (McCombs and Gay 1988; Parks and Kennedy 2007) or ethnicity (Glock and Krolak-Schwerdt 2013; van den Bergh et al. 2010). Such effects have been demonstrated in preservice teachers (Glock and Krolak-Schwerdt 2013; Parks and Kennedy 2007) and experienced in-service teachers (Glock et al. 2013; Parks and Kennedy 2007). The question arises to what extent the bias in the decision-making process is associated with (in)accuracy of decisions. To this extent, Jussim (2005) argued that if stereotypes accurately reflect certain differences between social groups (e.g., Jussim et al. 1996), bias would not necessarily lead to inaccuracy and could even enhance accuracy (Jussim 2005; Lee et al. 2013).

Most educational research concerning the assessment competence of teachers has focused on student characteristics such as ability and achievement, gender, and SES (Byrnes and Miller 2007; Demaray and Elliott 1998; Südkamp et al. 2012). Few, however, have applied insights from the field of social judgment formation and decision-making to study the way in which teachers select, use, and integrate student information into judgments regarding the future educational pathways of their students. Consequently, there are few theoretical explanations for judgment errors and incorrect decisions. In order to improve teachers' assessment competencies by training, it is necessary to develop theoretical explanations for judgment errors and to specify conditions of comparatively low or high assessment competence.

The findings reported in this chapter are from studies, which were conducted in Germany and in Luxembourg. As described above, in the educational systems of both countries, students are grouped into three school tracks based on teachers' assessments. However, some differences exist between the two countries in the formal procedures that teachers use to arrive at their decisions. In Germany, the teacher acts as an individual decision-maker such that the teacher's task is to recommend or decide on a secondary school track. In Luxembourg, a group ("council of orientation"), generally comprising of primary and secondary school teachers and school inspectors, makes tracking decisions, whereby it is the task of the primary school teacher to provide the group with information on the educational progress of his/her students before the group makes the tracking decisions. This implies a difference between the tasks of the teachers in terms of accountability for the decision, because

individual decisions are associated with higher accountability than group decisions (Lerner and Tetlock 1999). Taking into account teachers' tracking decisions in both Germany and Luxembourg broadens the scope of the present research by considering the nature of the task (individual decision vs. group decision).

5.3 Theoretical Framework of the Studies

To explain conditions of assessment competence in the educational context, theories of social judgment formation are used. Such theories conceptualize a decision as the result of a cognitive process involving not only the search for information but also the application of (implicit) rules regarding the use of information (Fiske and Taylor 2010; see also Alexander in this volume). More specifically, a decision is considered the result of a cognitive process including the attention to storage and retrieval of information about a student.

Different theories have been put forward to describe these cognitive processes. One theory assumes that people collect information in a systematic way, weigh, and integrate these informational cues when making a decision. Decisions are based on deliberate information-integrating strategies (e.g., Brehmer 1994; Dawes and Corrigan 1974; Swets et al. 2000), with a focus on individual characteristics, and are assumed to result in less biased decisions. Another theory assumes less complex judgment processes whereby a judge relies on a minimum of critical cues to make a decision by the use of cognitive heuristics (e.g., Gigerenzer and Todd 1999) and available social stereotypes determine the judgment. Although such heuristic, stereotype-based decision processes (Fiske and Taylor 2010; Hoffrage and Reimer 2004) are highly cognitively economical and efficient, they may be more prone to bias (see also Shavelson et al. in this volume; Oser et al. in this volume).

Dual-process theories of social cognition provide a framework to integrate the two information processing strategies, whereby people will switch strategy in response to differing task demands and motivational drives. To this extent, it is important to define conditions in which each strategy is most likely to occur. Starting with the assumption that people do prefer least-effort processing, which gives priority to the stereotype-based strategy, the continuum model (Fiske et al. 1999; Fiske and Neuberg 1990) posits that people are inclined to use this strategy when person information is consistent and a social stereotype is activated (Gilbert and Hixon 1991). More specifically, in situations in which available information automatically activates a social stereotype, decisions are mostly determined by the most salient attribute (e.g., gender, SES, or ethnicity). Such attributes may affect the judgments as social stereotypes direct the information processing toward selective attention to and retrieval of stereotypical information. Hence, resulting judgments are based not only on observed characteristics but may also be biased by inferred information (see also Alexander in this volume). In this regard teachers' acquired knowledge about members of social groups (e.g., Fiske and Taylor 2010) facilitates, but also colors,

perception and judgment formation (Ferguson 2003; Macrae et al. 1996). Stereotype biases in teachers' judgments are most likely to occur when student behavior or academic achievement is consistent and confirms stereotypical expectations (e.g., Jussim and Harber 2005).

In contrast, information-integrating strategies are used when cues are inconsistent and difficult to comprehend or when a person is motivated and has the cognitive resources to engage in the elaborate processing of individual information (Ferreira et al. 2006; Fiske and Neuberg 1990). Such motivation may derive from considering the consequences of the judgment for the target person, the expectation that the person has to justify his/her decision to others, or from high internal judgment standards set by the judge himself (Fiske and Neuberg 1990; Tetlock and Lerner 1999). In other words, such motivation may result from increased accountability or from the relative importance of the consequences of decisions (Gollwitzer and Moskowitz 1996; Pendry and Macrae 1996; Tetlock 1983). Hence, unbiased judgments are more likely to occur when a person is motivated to develop an accurate impression of the target person and hence is willing to invest cognitive effort.

Research has supported the notion that motivation moderates the use of processing strategy (Fiske and Taylor 2010; Gollwitzer and Moskowitz 1996) and that there is a continuous shift from stereotype-based to information-integrating processing, where a fully information-integrating strategy is only used as a last resort. Figure 5.1 gives a graphical display of these assumptions.

In sum, the continuum model posits how motivational and cognitive dispositions of decision-makers interact with the social context (especially accountability) to shape individual judgment and decision-making.

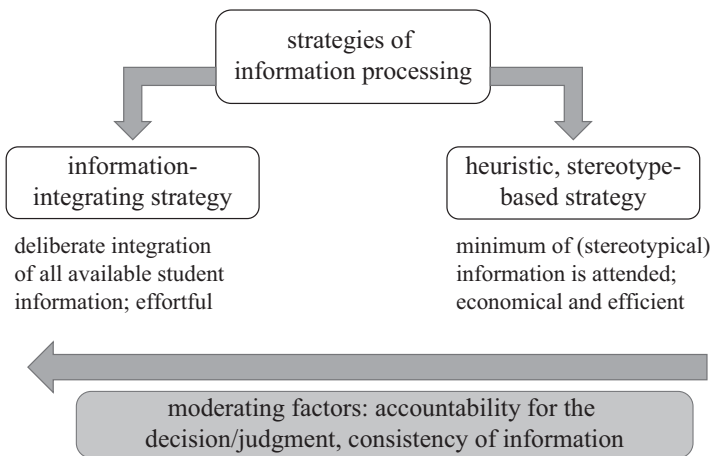


Fig. 5.1 Continuum model of judgment formation (Adapted from Fiske and Neuberg 1990; Tetlock 1992)

5.4 Teachers' Assessment Competence in Tracking Decisions: Results in Luxembourg and Germany

The research reported in this section consists of two complementing research lines. First, to investigate the extent to which students' academic achievement and social background influence actual teachers' tracking decisions, logistic regression models were used to analyze the relative importance of various predictors. The types of information considered in the field study were (1) teachers' tracking decision as the criterion and (2) social background of the student, (3) school grades at the end of primary school, and (4) teachers' informal assessments of students as predictors. In the Luxembourgish context, school grades at the end of primary school and those in the following 3 years in secondary school were available from a school monitoring project (see Sect. 5.4.1). As the progress of the students across the school transition into the following three school years could be followed, we were also able to analyze the prognostic validity of tracking decisions.

The second line of research investigated the cognitive processes underlying the tracking decision and was experimental in nature. Teachers of a sample drawn randomly from the field studies received student vignettes, which included the same types of information as the school monitoring project, whereby social background information and school grades were varied experimentally. In a series of experiments, participants' task was to read the information provided by the vignette, to select information cueing the tracking decision, and to decide upon an appropriate secondary school. The use of elaborated research techniques from cognitive psychology, such as the "eye-tracking", "Mouselab" technology, and "think aloud" method, allowed to determine which type(s) of information predominantly attracted attention and were retrieved from memory during the judgment process and how retrieved cues were combined into a decision.

This methodological approach allowed relating the experimental results to the corresponding findings from the field study. Consequently, the external validity of the experimental results could be assessed, and, vice versa, the formation of the tracking decision in the field studies could be causally explained.

5.4.1 Predictors of Teachers' Tracking Decisions

As mentioned above, in a first line of research, we examined to what extent students' achievement and social background influence the tracking decisions in the Luxembourgish school context. The collection of the data involving teachers' assessments of (real) students in their classes was implemented in the context of the Luxembourgish Ministry of Education's school monitoring program (for details, see Martin et al. 2015). In the school monitoring program, a database was created

which contained data for the complete Luxembourgish school population. From this database, valid data on 2702 school tracking decisions for students, who were sixth graders¹ during the 2008–2009 school year, were available. These students were distributed across 199 classes from a total of 108 primary schools. The final tracking decisions were made by a council consisting of teachers and school inspectors (see Sect 5.2). Furthermore, the database contained the following variables for all students (for details, see Klapproth et al. 2013): (a) the average *school grades* in the sixth grade in the main subjects *French*, *German*, and *mathematics*, (b) *test scores from standardized scholastic achievement tests* at the end of the sixth grade (French, German, mathematics), (c) students' *nationality*, and (d) the *gender* of the student. Students' *working behavior* was rated by their teachers, using Likert scales. *SES* was assessed by the HISEI index (Ganzeboom et al. 1992).

As only a small number of students (6%) were recommended for the lowest school track, tracking decisions were treated as a dichotomous variable where decisions were made either in favor of the academic track or in favor of the vocational tracks. Logistic regression analyses were conducted with the six types of student information (i.e., average *school grades*, *test scores*, *working behavior*, *nationality*, *gender*, and *SES*) as predictors. Results are presented in Table 5.1.

Four major results were obtained. First, students' school grades in language subjects (French, German) were most predictive of tracking decisions. The large weight of school grades in French, particularly, may reflect the fact that French is one of the main languages in secondary school. Second, school grades were on average more predictive of the tracking decision than scores of standardized tests (except mathematics). This result may reflect the predominance of school grades during instruction and teaching; however, it could also be that teachers mistrust results of standardized tests, as these results only partially capture curricular contents. Third,

Table 5.1 Estimated model coefficients in the Luxembourgish school monitoring study (Klapproth et al. 2013)

Predictor	Coefficient
German	4.37*
French	10.42*
Mathematics	2.07*
Test German	2.23*
Test French	2.39*
Test mathematics	3.89*
Working behavior	0.68*
Migration	0.71*
HISEI	2.22*
Gender	0.97

Note: * $p < 0.05$

¹In Luxembourg, the tracking from primary to secondary school takes place at the end of the sixth grade in primary school when the students are about 12 years old. In most parts of Germany, the tracking from primary to secondary school takes place at the end of the fourth grade in primary school when the students are about 10 years old.

SES as assessed by the HISEI substantially affected tracking decisions. This finding shows that higher socioeconomic status corresponded with higher probability of receiving a decision toward the highest track, even when achievement variables were entered into the regression analysis. Finally, inequalities in tracking decisions were obtained with respect to students' nationality. Even after controlling for variation in achievement, a significant contribution of the immigrant background of students to teachers' tracking decisions was observed, whereby students not originating from Luxembourg had a lower probability to be assigned to the highest track than Luxembourgish students. Thus, although the standards of the Luxembourgish Ministry of Education (Thill 2001) specify that only indicators of students' level of achievement in primary school (i.e., school grades in the main curricular areas, scores of standardized scholastic achievement tests, as well as students' working behavior) ought to determine the tracking decision, tracking decisions were clearly influenced by the social background of students.

The database of the school monitoring program also provided continuation rates of the students after 3 years of schooling in Luxembourgish secondary school, as well as individual scores on standardized scholastic achievement tests at the beginning of ninth grade. These data were used to analyze the predictive validity of tracking decisions. It was assumed that a high retention rate of students within the track they had been assigned to would be indicative of a high predictive validity of the tracking decision in primary school. To this extent, the very high continuation rate of 97% would suggest high predictive validity of the tracking decisions. However, consideration of results of the achievement tests in grade nine limits such conclusions. In support of predictive validity, we expected that students assigned to the highest track should achieve higher test scores than students allocated to one of the lower tracks, whereby the distributions of test scores should be clearly different for the various school types. However, analyses of the test score distributions showed that there was a high degree of overlap between the distributions of test scores related to the different tracks. This finding suggests that it is not the predictive validity of the tracking decision per se, but rather the low permeability of the Luxembourgish secondary school that leads to high retention rates. Therefore, continuation rates did not allow for a clear inference on the predictive validity of tracking decisions, and further research is needed taking into account additional measures of predictive validity (Klapproth et al. 2013).

In a similar approach in the German educational system, 56 German primary school teachers in charge of a fourth grade class reported information about each of their students as well as the actual school track recommendation they gave to the student (Böhmer et al. 2017). Teacher reported student information including (a) the school grades in the main subjects German, mathematics, and science, (b) student learning and working behavior, (c) student's immigrant background, and (d) available parental support. Students' immigrant background was indicated by a non-German nationality of the student or at least one of his/her parents.

Two multilevel regression analyses – one for the decision between the highest and intermediate school track and one for the decision between the intermediate and lowest track – showed results similar to those observed in the Luxembourg. Students'

school grades were the strongest predictors of highest school track recommendation, with odds ratios (OR) ranging from 6.4 for the grade in science to 13.8 for the grade in German. Working and learning behavior (OR = 3.1) and available parental support (OR = 2.0) explained additional variance in the school track recommendations, but not immigrant background. Similar results were observed for the decision between the intermediate and lowest school track. Interestingly, the same pattern of results also emerged if teachers did not make decisions for their actual students but based on 24 fictitious student descriptions including the same information as reported for the actual students.

In regard to the influence of social background variables on teachers' school track decisions, Böhmer et al. (2017) showed a significant influence of the parental support a teacher assumes to be available at the student's home. Although parental support is related to students' academic achievement (Jeynes 2005), it is also linked to the parental socioeconomic status (Rumberger et al. 1990) and immigrant background (Aldous 2006). As available parental support is not directly observable in the classroom, a teacher might derive assumptions about the parental support from the parents' nationality (e.g., insufficient German language skills) and parental profession (e.g., lower education and less financial resources). In this vein, disparities in students' social background may influence teachers' school track decisions, mediated by assumptions based on teachers' subjective knowledge about language skills and available support at home.

In sum, the studies in the Luxembourgish and German educational system pointed out that the social background of students influences their chances for a high-level academic career, as social disparities are reflected in teachers' secondary school track decisions. In regard to the observed risk of bias in tracking decisions, the question arises on how to explain the influence of social background information and, hence, social stereotypes, on decision processes by the use of dual-process theories of social cognition.

5.5 Experimental Studies on Teachers' Information Processing of Student Information

The studies described above as well as several large-scale studies (e.g., Burton et al. 2007; Stubbe and Bos 2008) have shown that teachers consider nonacademic information, especially information on students' socioeconomic and immigrant background, when judging student achievement and academic potential. In the framework of social cognition theories, the influence of socioeconomic and immigrant background information implies that teachers refer to stereotypes regarding students with different socioeconomic and immigrant background, which could lead to differential expectations about the students' academic achievement. Stereotypes are supposed to develop through experience with social groups (Stangor and Schaller 1996); thus, teachers might develop stereotypes by repeated observations of

immigrant or lower-class students performing worse in comparison to their classmates. In this line of reasoning, novice and preservice teachers should not necessarily have formed corresponding stereotypes.

5.5.1 Development of Stereotypes for Different Groups of Students

A set of studies investigated whether student teachers hold differential stereotypes for different groups of students. Adapting the design of Hofer's (1981) classical study on student stereotypes, Hörstermann and Krolak-Schwerdt (2012) asked Luxembourgish preservice teachers to make a free description of typical students they encountered during their prior teaching experience (ranging from several weeks to less than 1 year). Results of a cluster analysis showed ten distinct student stereotypes, which were described rather consensually by the student teachers. Beside academic achievement ("ideal student"), the student stereotypes covered a wide range of additional social ("clown in class," "outsider") and behavioral ("hyperactive student") attributes of students. In sum, the study pointed out that even preservice teachers already hold a differentiated set of student stereotypes. Thus, student stereotypes develop rather fast through minimum teaching experience or are already developed at the beginning of teacher education (e.g., through social learning processes or preservice teachers' own experiences as students).

From this we can conclude that student stereotypes are present at the beginning of teachers' professional career as sources of judgment bias. Therefore, interventions to reduce bias in student assessment could already be included into initial teacher education, offering an easier and more widespread access than later in-service teacher education.

5.5.2 How Teachers Search and Process Student Information

Research of teachers' assessment of students has mainly adopted an output-oriented approach (i.e., how information about students influences teachers' judgments and leads to systematic bias in teacher judgments). However, some studies have focused on a more detailed analysis of teachers' information search processes. More specifically, these studies have considered how teachers gather information about a student to be used in later judgments of student academic potential (Böhmer et al. 2012, 2015). In an experimental design using the "Mouselab" technology, German primary school teachers were presented with an information matrix of 25 pieces of information on a computer screen (see Fig. 5.2). Information was presented as labeled information fields (e.g., autonomy), which could be opened by one mouse-click to read the corresponding piece of information.

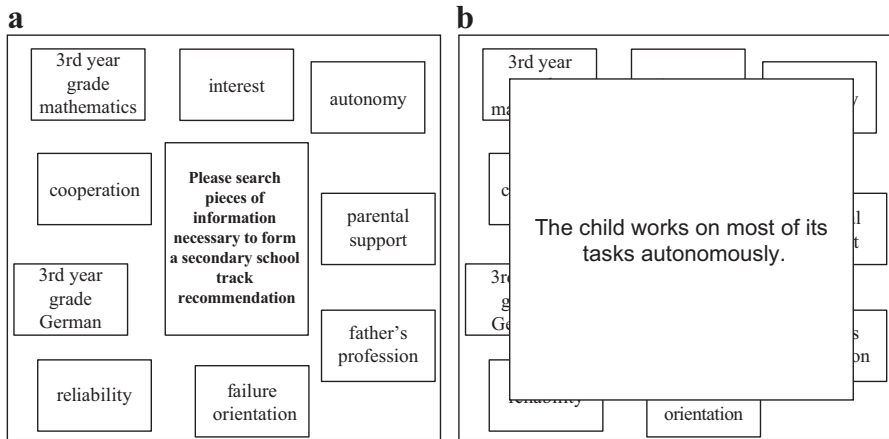


Fig. 5.2 Schematic illustration of (a) “the information matrix,” and (b) “piece of information” displayed

Information fields encompassed information on school grades; working, learning, and social behavior; as well as social background (i.e., parental support and immigrant background). In total, four information matrices were presented sequentially, and teachers were instructed to search for all pieces of information they required in order to decide on an adequate secondary school track for each student. Before each information matrix, teachers were informed about the student’s current school grades in the main subjects German, mathematics, and science. For two of the students (consistent student profiles), these school grades unambiguously reflected either high or low academic achievement, thus clearly implying a decision for a high or low secondary school track. For the other two students (inconsistent student profiles), the school grades were mixed, showing medium to high academic achievement or low to medium academic achievement, respectively. Hence, different pieces of information supported decisions for different secondary school tracks.

Analyses of the teachers’ search behavior (frequencies) showed that information on prior school grades as well as working and learning behavior was most often searched by the teachers. Moreover, these information cues were searched at the very beginning of the information search process. Comparing consistent and inconsistent student descriptions, results showed a higher total number of information searched for inconsistent students, whereby this difference holds true for nearly every piece of information. Highest effect sizes between consistent and inconsistent student profiles were observed for information on working and learning behavior as well as on parental support. Regarding the sequence of information searched, no differences were shown between consistent and inconsistent student profiles, indicating that teachers did not alter their information search sequence depending on the consistency, but rather searched along the same sequence more intensively for students with inconsistent profiles.

In sum, the results demonstrate a situational component of teachers' competence. Depending on the consistency of the student profile and thus difficulty of the decision, teachers adapted the extent of their information search and, hence, cognitive capacity invested, indicating an adaptive switch of information processing strategies. However, results also showed possible sources for bias on the process level. More specifically, for students with inconsistent profiles information on parental support was more often searched. As parental support may be related to parents' socioeconomic and educational background, and especially if teachers use parents' socioeconomic and educational background as a proxy for available support, decisions for students with inconsistent academic profiles might be prone to bias associated with the student's social background.

To investigate the role of teacher education or professional experience, information processing strategies of novice and laymen were compared to expert teachers. To this extent, the exact same information search task was presented to German preservice teachers (Böhmer et al. 2012). Results showed that preservice teachers – similar to expert teachers – adapted the extent of their information search to the consistency of the student profiles, whereby the especially high effect size for information on parental support was also observed. This finding indicates that preservice teachers are already able to adaptively switch their information processing strategies. Furthermore, a similar proneness to judgment bias for students with inconsistent academic profiles can be implied. In contrast to expert teachers, preservice teachers searched for more irrelevant and social background information, independent of the consistency of student profiles. Hence, preservice teachers seem to be less able of differentiating between achievement-related and other information. This indifferent pattern of information search might lead to dilution effects in judgment (i.e., achievement-related information is blurred by less valid information).

A further investigation of the ability to differentiate between more and less relevant information in information processing was conducted by using the eye-tracking method (Hörstermann et al. 2017). In this study, student profiles were presented to Luxembourgish laymen with no teaching experience. Student profiles included information on social background, school grades, standardized test scores, working behavior, and class repetitions. In order to investigate a spatial primacy effect, half of the participants were presented profiles in which social background information was positioned at the top-left position of the profile and grade information was positioned on the top-right position. For the other half of the participants, the positioning was reversed. The content of the profiles and the position of the other pieces of information were kept constant. Assuming that participants' focus of attention should be determined by the importance of the information, not by its position, the positioning of information should not influence the information processing when deciding about an adequate secondary school track. However, in accordance with the left-to-right and top-to-bottom orientation in Western European languages, it was hypothesized that the information positioned at the top-left position would draw more initial attention and more attention in total as well.

Investigating the process of information processing in terms of attention directed to the information, the eye-tracking data showed a spatial priming effect: social

background information in top-left position drew more initial attention and lowered the total attention directed to grade information. Thus, information positioning might influence the focus of attention, possibly increasing attention to especially bias-prone information (e.g., social background information). As teachers in their daily practice deal with a variety of documents providing information about students (e.g., files, certificates, and report cards), and these documents usually follow a common structure, in which personal information (e.g., student's name, age, and nationality) is presented at the beginning of the document, these documents might emphasize attention to social background information. Although the study does not allow conclusions on teachers, it identifies spatial primacy effects as possible bias source in information processing.

A further set of studies was concerned with the impact of social background information on teachers' information processing. Glock and Krolak-Schwerdt (2014) investigated the effect of social background information on attention to and recall of student information in two studies. In the first study, participants received the same student information where half of the participants received additional social background information to activate a stereotype. Participants' task was to recall and write down the information they could remember. Higher recall and higher intrusion rates were found in the condition of stereotype activation. Thus, social background information as a stereotype affected structuring and storing of the student information. In the second study, a self-paced reading time method was used to analyze in how far attention is affected by social background information. Reading times were faster for stereotype-related information than for non-stereotypical information, thus showing that stereotype-related information facilitated comprehension of the information. Glock et al. (2013) investigated the role of both social background information and profile (in)consistency on teachers' judgments by the use of the think aloud method. Participants were given student descriptions varying in profile consistency, where half of the participants received additional social background information. Participants' task was to think aloud while reading and judging the different student descriptions. Think aloud data indicated more careful processing of all information both for students presented with social background information and students with inconsistent profiles. Taken together, these studies clearly demonstrated that social background information and profile (in) consistency affected phases of teachers' information processing, comprising of attention, storage, recall, and putting cues together for a judgment.

5.6 Motivation Matters: Accountability Affects Teachers' Tracking Decisions

Results on teachers' search and information processing outlined above pointed to the relevance of (in)consistency of student information, which seemed to activate different information processing strategies. According to the theoretical framework

of the present studies, these findings may be explained by the motivation of the decision-maker and, more specifically, by the accountability for the decision at hand. Judgments of teachers who are confronted with tasks for which they are highly accountable should be less affected by stereotype biases and other nonachievement-related information than judgment for which teachers perceive low accountability. Another series of experimental studies were dedicated to investigate the impact of accountability on teachers' judgments. Accountability should decrease when teachers have no need to justify their decisions to others, for instance, to the parents, and when the decision is of no particular importance for the teacher or for the student. In this case, it is likely teachers use a stereotype-based processing strategy. In contrast, teachers who feel more accountable (i.e., when they have to justify their decisions, when they are personally responsible for the judgment, and when the decisions are highly important for the students' lives) should use an information-integrating strategy employing the rule that achievement-related cues of the individual student provide the best information for making an accurate decision.

Glock et al. (2012) investigated these assumptions. In this study, Luxembourgish primary school teachers were presented with vignettes, describing primary school students using the same types of information as in the Luxembourgish school monitoring study (see Sect. 5.4.1), that is: (a) *school grades* in the main subjects, (b) *test scores* on the standardized school achievement tests conducted in the main subjects, (c) *working behavior*, (d) *nationality*, (e) *gender*, and (f) *SES*. In addition, information on *social behavior* (i.e., information about the students' behavior during instruction and school recess) was added as another indicator of nonachievement-related information. Teachers were asked to decide which secondary school track each student should attend.

To vary levels of accountability, three different instructions were presented to different groups of teachers. In the high-accountability group, teachers were asked to imagine that they were solely responsible for the tracking decisions and that these decisions would influence the future educational and occupational careers of the students. In the low-accountability group, teachers were asked to imagine a situation in which a colleague would ask them for advice concerning the tracking decisions for the students of his/her class and that they were just required to provide their opinion without commitment. In the third group ("council instruction"), teachers were asked to prepare the tracking decisions for the council and were informed that, in accordance with the actual procedure in Luxembourg, the final tracking decisions would be made by the council.

To analyze the tracking decisions, a multiple logistic regression analysis for each experimental condition was used. Results are presented in Table 5.2.

For the high-accountability group, the achievement-related cues (i.e., school grades, test scores, and working behavior) predicted teachers' tracking decisions, whereby higher scores were associated with a higher chance of choosing the highest track. In the low-accountability group, working behavior was not considered as a significant cue. However, only in the low-accountability group teachers' tracking decisions were additionally based on the nationality of the student as nonachievement-related cue. More specifically, in the low-accountability group and given similar

academic achievements of students, students without immigration background had approximately a three times higher chance of being recommended to the highest track than students with immigration background. For the council condition, the study provided mixed results. It was expected that teachers who made their decisions in the council and who had no need to justify their decisions to others would exhibit stereotype-based processing. The regression analysis for the tracking decisions indicated on the one hand that teachers clearly relied on school grades and test scores as achievement-related cues but, on the other hand, that they did not take working behavior as another achievement-related cue into account. An additional manipulation check on perceived accountability of each set of instructions confirmed a medium level of perceived accountability between high and low accountability and received an average score between the low- and high-accountability conditions. A replication of the experiment provided the same results on teachers' tracking decisions (Glock et al. 2012).

These studies showed that teachers use stereotype-based or information-integrating processing strategies depending on their accountability for the decision with the council condition between the more extreme poles of high versus low accountability. However, the studies did not provide evidence for the hypothesis that teachers should be able to switch between processing strategies depending on their accountability as manipulation of accountability was varied in a between-subjects design. Therefore, Krolak-Schwerdt et al. conducted two studies in which the experimental manipulation of accountability was varied in a within-subjects design. Primary school teachers received vignettes of students, which comprised of either consistent or inconsistent student information both under high- and low-accountability instructions. Results indicated that teachers applied a stereotype-based processing strategy when they were confronted with consistent student cases under low accountability, while inconsistent student cases as well as high accountability led to an information-integrating strategy.

Taken together, the findings of these studies confirm the hypotheses concerning the role of motivation derived from the theoretical framework of this chapter.

Table 5.2 Odds ratios for the seven cues in each accountability condition (Glock et al. 2012)

Predictor	Low accountability	Council	High accountability
School grades	85.18*	31.86*	50.67*
Test scores	14.01*	12.32*	10.26*
Nationality	3.21*	1.84	1.68
SES	1.20	1.54	1.53
Working behavior	2.42	2.01	2.81*
Social behavior	1.53	1.38	1.50
Gender	2.04	2.11	1.56

Note: * $p < 0.05$

5.7 Bias and Accuracy

The research discussed so far has demonstrated that teacher judgments and associated decisions may be biased by stereotypical expectations associated with different groups of students. However, the extent to which this bias reflects objective reality or inaccurately favors or hinders one group over the other remains unclear (Jussim 2005; Lee et al. 2013). One problem in this line of research is that it has been challenging to develop materials and criteria that can be used to reliably and validly assess accuracy (Jussim 2005). In our research, we therefore aimed to define and validate a criterion to judge the accuracy of tracking decisions. In a next step, we then applied this criterion to investigate teacher decision-making accuracy and in particular the extent to which accuracy would be affected by bias. Using data from the Luxembourgish school monitoring project (see Sect. 5.4.1), where school grades, test results, and actual tracking decisions of a student cohort were known (Klapproth et al. 2012), and including only the most salient academic predictors of school tracking decisions (Klapproth et al. 2013), the criterion was based on the likelihood of an observed achievement score pattern in relation to achievement score distributions in different school tracks (Pit-ten Cate and Hörstermann 2012).

Fig. 5.3 Schematic illustration of the relative fit of a typical student for the highest track

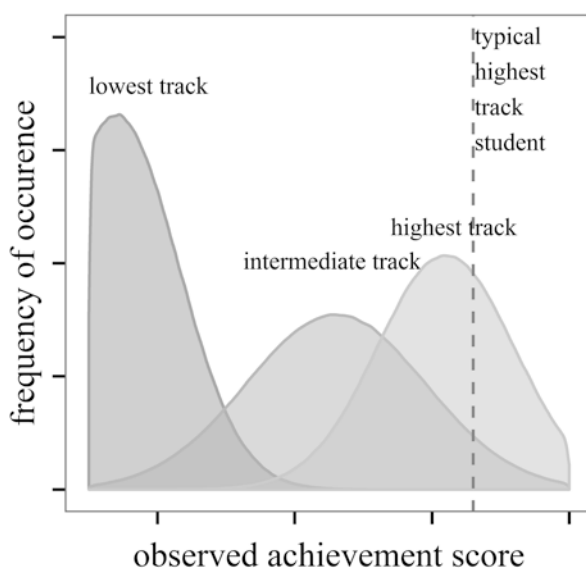


Figure 5.3 gives an illustration of relative fit of a fictitious student typical for the highest school track.

This way we could differentiate between students with consistent (clear fit to a specific school track) and mixed (fitting two school tracks equally well) academic profiles. The criterion's predictive validity was first tested by the match between criterion-based and actual tracking decisions. Cross tab analyses indicated that there

was a 90% match between the criterion-based and actual tracking decisions. We then investigated the extent to which the criterion could predict student performance in the third year of secondary school. More specifically, we assessed the relationship between students' test scores (Fischbach et al. 2014) and the criterion-based decisions. Results showed that test scores of students with consistent profiles reflected academic differences at the time of the tracking decision. For the highest track, "correctly" tracked students generally outperformed students "incorrectly" downwardly tracked students. For the other tracks, upward or downward tracking did not result in significant performance differences. For students with mixed profiles, a similar pattern was observed for the students with lower achievement scores, whereby students referred to the higher of the two matching tracks outperformed students with similar academic profiles referred to the lower of the two matching tracks (Pit-ten Cate et al. 2015). These results demonstrated that incorrect tracking downward disadvantages high-performing students with consistent profiles and average or below-achieving student with mixed profiles. We therefore considered that for students with mixed profiles, decisions to refer students to the higher track would be considered accurate.

In a second study, we applied the criterion to investigate teacher decision-making accuracy and in particular the extent to which accuracy would be affected by stereotype bias. For this study, we created student vignettes using actual (anonymized) student data for correctly orientated students (i.e., reflecting a match between criterion-based and actual decisions and for students with mixed profiles and actual decision to the higher of the two tracks). The vignettes presented demographic information, academic achievement data, working and learning habits, class repetitions, and parental track preference. We then presented vignettes (equally representing each school track and systematically varying the consistency of the student's academic profile and student immigrant background) to experienced primary school teachers with the question to provide a tracking decision for each. Results of a repeated measure ANOVA (analysis of variance) revealed that teachers were more accurate in their decisions for students with consistent academic profiles than for students with mixed profiles. In addition, their decisions for students without an immigrant background were more accurate than for students with an immigrant background, but only for students with consistent profiles (Pit-ten Cate et al. 2015). From the combined results, we can conclude the criterion is quite robust and can be considered a valid measure to which individual teacher judgment can be compared. Although teachers were generally quite accurate in their decisions, a stereotype bias associated with the student's immigrant background bias existed; whereby decisions for minority students were less accurate, demonstrating that bias was associated with inaccuracy of judgments.

5.8 Training and Intervention Studies

So far, we have presented studies demonstrating that teachers' tracking decisions may be affected by stereotype bias, especially in situations in which teachers do not feel highly accountable for their decisions. Furthermore, the results of the presented studies have supported the notion that the different processes involved in the judgment formation (i.e., attention, search for information, retrieval of information, and the application of different information processing strategies) provide insight in the underlying mechanism that may explain conditions under which bias is most likely to occur and why. Different strategies can be employed to reduce such bias (Pit-ten Cate et al. 2014). For example, research has shown bias can be reduced by increasing the accountability (e.g., Krolak-Schwerdt et al. 2013). Another strategy involves providing theoretical knowledge on judgment formation and training under feedback conditions (Helmke et al. 2004). Yet a third strategy is to apply formal decisions rules on the weighted integration of information (Brehmer 1994; Swets et al. 2000). Such rules contain a defined weighting of the key information that teachers use in the decision-making process. For example, before making tracking decisions, a teacher, in accordance with the stipulation that such decisions should primarily be based on academic achievement, may define a weight of 75% to school grades and 25% to results of a standardized achievement test, with a weight of 0% for all other variables. However, in their actual judgment formation, teachers may deviate from these preset rules by also considering other student variables (e.g., working behavior or background information). Providing feedback about the deviation between preset and actual consideration (i.e., weight) of (non)achievement-related information provides insight and raises awareness of the potential effect of stereotype-based expectations on teachers' judgments. Using these insights, we developed and evaluated different intervention modules, aimed to reduce bias in decision-making and hence increase the accuracy of the resulting decisions. Using an experimental pre-post design, we investigated the effect of accountability and training on the reduction of bias in tracking decisions, especially bias against students with immigrant backgrounds.

In the first module, we increased accountability by instruction (Pit-ten Cate et al. 2016a, b). A manipulation check showed that teachers felt significantly more responsible for their tracking decisions after this instruction than before. For the two other modules, we delivered and evaluated separate workshops on theoretical knowledge and the application of decision rules to experienced teachers (Pit-ten Cate et al. 2013, 2016a, b). In the first workshop, we introduced theoretical models of judgment formation and gave teachers feedback concerning their tracking decisions and student-related inferences. We presented an overview of theories regarding decision-making and accuracy and discussed factors and conditions affecting

the application of different information processing strategies. Then, teachers inter-actively developed strategies for making tracking decisions, which they applied under feedback conditions. The second workshop focused on the application of formal rules on the weighted integration of student information. We devised a computerized training module comprising of four different stages, in which teachers first rated the relevance of different student attributes for the tracking decision. Subsequently, individual decision rules were computed, reflecting an optimized prediction in accordance with the teacher's intended decision-making strategy. Then teachers made tracking decisions after which they received immediate feedback about the concordance between the predicted and actual decision. Teachers were asked to make tracking decisions for students based on vignettes before and after the intervention.

Results of a repeated measures analysis of variance showed that although teachers' tracking decisions were generally quite accurate, they made more accurate decisions for student without immigrant background than for students with immigrant background. Furthermore, teachers' decisions were more accurate after the intervention than before the intervention, whereby teachers' tracking decisions became only more accurate for students with immigrant background. More specifically, before the intervention, teachers' tracking decisions were less accurate for students with immigrant background than for students without immigrant background, whereas after the intervention, this bias disappeared. Results were independent of the type of workshop. These results showed that although teachers were generally good decision-makers, disadvantages for students with immigrant backgrounds in teachers' tracking decisions could be reduced by training. In line with the intention of the workshops, the disproportionately high rate of decision errors for students with immigrant background observed before the workshops was eliminated and in correspondence with error rates for students without immigrant background. Thus, overcoming or increasing awareness of effects of stereotypical expectations on judgments, either via increased accountability, theoretical knowledge, or the systematic application of formal decisions rules, can successfully increase diagnostic competence by reducing differences in decisions for students with and without immigrant background.

5.9 Conclusions

Over the last decennia, several (large scale) studies have reported bias in teacher judgments, which have resulted in inequalities in educational systems (Baumert et al. 2001; Bos et al. 2004; OECD 2010). Especially students with immigrant background and from low-income families fare less well and have fewer educational opportunities. Traditionally, research on teachers' diagnostic competence has focused on observable student or judgment outcomes, and only few have considered the decision-making processes or provided theoretical explanations for judgment errors in educational contexts. Theoretical explanations for judgment errors and

especially the specification of conditions that may reduce bias and hence increase judgment accuracy are an important prerequisite to improve teachers' diagnostic competence by training. Therefore, the studies presented in this chapter have considered the extent to which insights from social psychology and social cognition can facilitate understanding of teachers' information processing and judgment formation. In a series of studies, conducted within several externally funded projects in both Germany and Luxembourg, we have been able to demonstrate a stereotype bias in teachers' decisions (e.g., Glock et al. 2010, 2013, 2015; Klapproth et al. 2012). The influence of stereotype bias is observable in different stages of the information processing. More specifically, we have demonstrated that nonachievement-related student information affects the search for and attention to information (e.g., Böhmer et al. 2012, 2015; Hörstermann et al. 2017), the recall of information (e.g., Glock et al. 2015; Hörstermann et al. 2017), and the accuracy of resulting judgments (Pit-ten Cate et al. 2015, 2016a, b). These findings were independent of experience (i.e., similar for pre- and in-service teachers) and country (i.e., similar for teachers in Germany and Luxembourg). Based on dual-process theories of judgment formation (Fiske et al. 1999; Fiske and Neuberg 1990), which have posited that individuals switch between stereotype-based and information-integrating processing strategies depending on the information availability, the (in)consistency of cues, and the person's motivation, we conducted studies to investigate the extent to which the (in) consistency of information or perceived accountability affects the teachers' information processing and judgments. Results indicated that with inconsistent student information (Glock and Krolak-Schwerdt 2013) and under conditions of high accountability (e.g., Glock et al. 2012; Krolak-Schwerdt et al. 2013; 2016a, b), teachers were more likely to apply elaborate information-integrating strategies, leading to a reduction of bias and increased accuracy of decisions. In a next step, intervention studies were conducted to test the effect of different strategies aimed to reduce bias in decision-making (Pit-ten Cate et al. 2014). Results indicated that bias can be reduced by increasing the accountability (Pit-ten Cate et al. 2016a, b), providing theoretical knowledge on judgment formation, and training under feedback conditions or the application of formal decisions rules on the weighted integration of information (Pit-ten Cate et al. 2013, 2016a, b).

Taken together, the different projects have provided substantial insight into the extent to which underlying mechanisms of information processing and judgment formation can explain repeated findings of stereotype bias in teachers' decisions regarding student achievement and potential. Another research line not reported in this chapter investigated the effect of (in)consistency of student information and accountability on teachers' school grades in secondary school (e.g., Krolak-Schwerdt et al. 2013) also using dual-process theories of judgment formation as a theoretical base. Results were comparable to those reported in this chapter on teachers' tracking decisions, which suggests that the reported findings correspond to more general principles of teachers' achievement judgments and diagnostic competence.

It should be noted that in general teachers are competent decision-makers (Pit-ten Cate et al. 2016a, b; Südkamp et al. 2012) but that there is still room for improve-

ment. In regard to tracking decisions, our results have demonstrated that by changing the situational context (i.e., under conditions of high accountability), the effect of bias reduces significantly. To this extent, it is interesting to note different levels of teacher accountability in tracking procedures in Germany, where the teacher's decision is either binding or can be perceived as a recommendation (Nölle et al. 2009; Ditton and Krüsken 2009), which is usually accepted by parents and students, as well as proposed changes in the tracking procedure in Luxembourg where, as of the 2017, the main responsibility for the tracking decision is referred to the teacher (Pit-ten Cate and Krolak-Schwerdt 2016).

On a theoretical level, the findings have resulted in the formulation of a process-based decision-making model. This is an adaptation of the dual-process model for application in the educational context. In accordance with the dual-process model, the adaptive diagnostic competency model (Böhmer et al. 2017) poses that teachers can adapt their information processing strategies in response to the situational demands. More specifically the model states the conditions under which teachers are likely to switch between heuristic processing of stereotype-based information and information-integrating processing of students' individual abilities and behaviors. This model provides a valuable addition to other models of professional and diagnostic competence (Baumert and Kunter 2006; Bruder et al. 2010) as it specifically focuses on the situational context in which the judgment formation takes place. The adaptive diagnostic competency model also fits within the theoretical framework of the working model on teacher assessment competence (Herppich et al. 2017), which defines diagnostic competence as context-specific, cognitive dispositions of achievement which enables teachers to master assessment requirements in different pedagogical situations and links knowledge structures to real-life performance in specific action situations. To this extent, the model can be used as an additional tool in the assessment and training of teachers' diagnostic competence.

Different modules have previously been developed to study teachers' diagnostic competence in relation to different group of students (e.g., Kaiser et al. 2013; Seidel and Prenzel 2007; Südkamp et al. 2008). For example, simulated classroom paradigms have been applied to assess teachers' competence (e.g., Kaiser et al. 2013; Südkamp et al. 2008) but could also be used as a training tool, whereby teachers could receive feedback concerning their classroom behaviors. Similarly, videotaped classroom situations have been used to assess teachers' competence (Seidel and Prenzel 2007) but also applied in training (Seidel et al. 2011).

An obvious next step following the formulation of the adaptive diagnostic competency model as a theoretical framework to integrate our findings as well as the conceptualization and evaluation of training modules is the development of an integrative instrument to measure the diagnostic competence of teachers according to the underlying theoretical model. For some specific aspects, measurement instruments have already been conceptualized. For example, the training module using prediction rules involves a powerful instrument to measure the extent to which an information-integrating strategy is used and which types of diagnostic cues are adequately weighted while forming (highly consequential) judgments or decisions. Teachers' insight into the role of stereotypes and different types of decision-making

might be quite easily measured by a conventional knowledge test. More challenging is the measurement of the dynamic, situation-dependent part of the model. In this respect, new approaches to measure teachers' capability to switch between different processing strategies depending on the situational context and accountability demands are called for.

Our current insights could extend existing training models by the development of theoretically driven and research-based training for both primary and secondary school teachers. The intervention modules reported in this chapter were evaluated within an in-service teacher training framework and are therefore readily transferable to continuous education programs of experienced teachers. Given the findings for preservice teachers, especially the observed existence of student stereotypes and the increased attention to nonacademic (e.g., social background) information, an integration of the evaluated intervention modules into university teacher training appears to be promising. The inclusion of theoretical models of judgment formation in teacher training programs may foster preservice teachers' understanding of the nature of stereotypes, the processes causing stereotypes to influence judgments, and the flexibility of judgment formation processes. Thus, preservice teachers learn that stereotypes are integral part of human cognitions but can be overcome by avoiding stereotype-based information processing in important decision situations. Furthermore, providing an overview about the validity of different pieces of student information in the assessment of student achievement can improve preservice teachers' ability to separate between more and less valid information, concentrating their focus of attention to highly valid, achievement-related information. To this extent, training could incorporate a practical component in the application of prediction rules that can be applied under feedback conditions. This would facilitate a reflective and systematic approach to judgment formation. Furthermore, training modules focusing on increased accountability may provide techniques to preservice teachers to actively regulate their choice of information processing strategy. By being aware that the situational context (i.e., the level of accountability) influences information processing, preservice teachers learn deliberate attempts to increase accountability. For example, perceived accountability can be increased by considering the consequences of the decision at hand, thus willingly promoting information-integrating processes in highly consequential decision settings (e.g., school track decisions). Eventually, these interventions might result in self-instruction techniques that preservice teachers can transfer to their later in-service activity. Intervention modules providing feedback on the judgment formation process can serve as a formative assessment throughout the outlined student teacher training, offering feedback on the relative impact of different pieces of student information on student teachers' judgments, thus allowing preservice teachers to continually evaluate their success in avoiding less valid and stereotypical information to influence their judgments.

Acknowledgments The research reported in this chapter was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) and the Research Council of Luxembourg (Fonds National de la Recherche, FNR) grants KR 2162/4-1, INTER/DFG/09/01, INTER/DFG/11/03, C08/LM/02, and C10/LM/784116.

References

- Aldous, J. (2006). Family, ethnicity, and immigrant youths' educational achievements. *Journal of Family Issues*, 27, 1633–1667.
- Alpert, B., & Bechar, S. (2008). School organisational efforts in search for alternatives to ability grouping. *Teaching and Teacher Education*, 24, 1599–1612.
- Ansalone, G. (2001). Schooling, tracking, and inequality. *Journal of Children and Poverty*, 7, 33–47.
- Ansalone, G., & Biafora, F. (2004). Elementary school teachers' perceptions and attitudes to the educational structure of tracking. *Education*, 125, 249–260.
- Bamberg, M., Barthelemy, M., Bertemes, J., Besch, E., Boehm, B., Brunner, M., et al. (2010). *PISA 2009: Nationaler Bericht Luxemburg* [PISA 2009: National report for Luxembourg]. Luxembourg: MENFP-SCRIPT & University of Luxembourg; EMACS.
- Baumert, J., & Kunter, M. (2006). Stichwort : Professionelle Kompetenz von Lehrkräften [Keyword: Professional competencies of teachers]. *Zeitschrift für Erziehungswissenschaft*, 9, 469–520.
- Baumert, J., & Schümer, G. (2002). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb im nationalen Vergleich [Family background, selection and achievement: The German experience]. In *PISA 2000 — Die Länder der Bundesrepublik Deutschland im Vergleich* (Vol. 5, pp. 159–202). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (2001). *PISA 2000 : Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* [A international comparison of basic competencies of students]. Opladen: Leske & Budrich.
- Bertemes, J., Boehm, B., Brunner, M., Dierendock, C., Fischbach, A., Gamo, S., et al. (2013). *PISA 2012: Nationaler Bericht Luxemburg* [PISA 2012: National report for Luxembourg] (MENFP-SCRIPT and EMACS, Ed.). Luxembourg: MENFP and University of Luxembourg.
- Böhmer, I., Gräsel, C., Hörstermann, T., & Krolak-Schwerdt, S. (2012). Die Informationssuche bei der Erstellung der Übergangsempfehlung – Die Rolle von Fallkonsistenz und Expertise [Information search in decisions on school tracking recommendations – The influence of case consistency and expertise]. *Unterrichtswissenschaft*, 40, 140–155.
- Böhmer, I., Hörstermann, T., Gräsel, C., Krolak-Schwerdt, S., & Glock, S. (2015). Eine Analyse der Informationssuche bei der Erstellung der Übergangsempfehlung. Welcher Urteilsregel folgen Lehrkräfte? [An analysis of information search in the process of making school tracking decisions: Which judgment rule do teachers apply?]. *Journal for Educational Research Online*, 7, 59–81.
- Böhmer, I., Gräsel, C., Krolak-Schwerdt, S., Hörstermann, T., & Glock, S. (2017). Teachers' school tracking decisions. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments* (pp. 131–147). Berlin: Springer.
- Bos, W., Voss, A., Lankes, E.-M., Schwippert, K., Thiel, O., & Valtin, R. (2004). Schullaufbahneempfehlungen von Lehrkräften für Kinder am Ende der vierten Jahrgangsstufe [Teachers' school track recommendations at the end of 4th grade primary school]. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin, & G. Walthert (Eds.), *IGLU- Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* (pp. 191–228). Münster: Waxmann.
- Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica*, 87, 137–154.
- Bruder, S., Klug, J., Hertel, S., & Schmitz, B. (2010). Modellierung der Beratungskompetenz von Lehrkräften. Projekt Beratungskompetenz [Modeling teachers' counseling competence. Project counseling competence]. *Lehrerbildung Auf Dem Prüfstand, Beiheft*, 56, 173–193.
- Burton, R., Reichert, M., Brunner, M., Keller, U., Böhm, B., & Martin, R. (2007). Migrationshintergrund und sozioökonomischer Hintergrund der Schülerinnen und Schüler [Immigrant- and socioeconomic background of students]. In MENFP and EMACS (Ed.), *PISA*

- 2006 – *Nationaler Bericht Luxemburg* (pp. 32–45). Luxembourg: MENPF and University of Luxembourg.
- Byrnes, J. P., & Miller, D. C. (2007). The relative importance of predictors of math and science achievement: An opportunity–propensity analysis. *Contemporary Educational Psychology, 32*, 599–629.
- Dauber, S. L., Alexander, K. L., & Entwisle, D. R. (1996). Tracking and transitions through the middle grades: Channeling educational trajectories. *Sociology of Education, 69*, 290.
- Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95–106.
- Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*, 8–24.
- Ditton, H., & Krüsken, J. (2009). Bildungslaufbahnen im differenzierten Schulsystem – Entwicklungsverläufe von Laufbahneempfehlungen und Bildungsaspirationen in der Grundschulzeit. [Educational careers in a tracked school system – Development of teacher recommendations and educational aspirations over the elementary school years]. *Zeitschrift für Erziehungswissenschaft, 12*(Sonderheft 12), 74–102.
- Dustmann, C. (2004). Parental background, secondary school track choice, and wages. *Oxford Economic Papers, 56*, 209–230.
- Ekstrom, R. B., Goertz, M. E., & Rock, D. A. (1988). *Education and American youth*. Philadelphia: Falmer.
- Ferguson, R. F. (2003). Teachers' perceptions and expectations and the black-white test score gap. *Urban Education, 38*, 1–49.
- Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgment and decision making. *Journal of Personality and Social Psychology, 91*, 797–813.
- Fischbach, A., Ugen, S., & Martin, R. (2014). *ÉpStan technical report*. Luxembourg: University of Luxembourg.
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). New York: Academic.
- Fiske, S. T., & Taylor, S. E. (2010). *Social cognition: From brains to culture*. New York: McGraw-Hill.
- Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model. Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 231–254). New York: Guilford Press.
- Gamoran, A. (1992). The variable effects of high school tracking. *American Sociological Review, 57*, 812–828.
- Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard socio-economic index of occupational status. *Social Science Research, 21*, 1–56.
- Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 3–34). Oxford: Oxford University Press.
- Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypical beliefs. *Journal of Personality and Social Psychology, 60*, 509–517.
- Glock, S., & Krolak-Schwerdt, S. (2013). Does nationality matter? The impact of stereotypical expectations on student teachers' judgments. *Social Psychology of Education, 16*, 111–127.
- Glock, S., & Krolak-Schwerdt, S. (2014). Stereotype activation versus application: How teachers process and judge information about students from ethnic minorities and with low socioeconomic background. *Social Psychology of Education, 17*, 589–607.
- Glock, S., Krolak-Schwerdt, S., Zöllner, I., & Martin, R. (2010). Grundschemphelungen in Luxemburg – gleiche Chance für alle? [Tracking decisions in Luxembourg – equal chances for

- all?]. In F. Petermann & U. Koglin (Eds.), *Erklären, Entscheiden, Planen. 47. Kongress der Deutschen Gesellschaft für Psychologie. Abstracts* (p. 46). Lengerich: Pabst Science.
- Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2012). Improving teachers' judgments: Accountability affects teachers' tracking decisions. *International Journal of Technology and Inclusive Education, 1*, 89–98.
- Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2013). Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Social Psychology of Education, 16*, 555–573.
- Glock, S., Krolak-Schwerdt, S., & Pit-ten Cate, I. M. (2015). Are school placement recommendations accurate? The effect of students' ethnicity on teachers' judgments and recognition memory. *European Journal of Psychology of Education, 30*, 169–188.
- Gollwitzer, P. M., & Moskowitz, G. B. (1996). Goal effects on action and cognition. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 361–399). New York: Guilford Press.
- Haller, E. J. (1985). Pupil race and elementary school ability grouping: Are teachers biased against black children? *American Educational Research Journal, 22*, 465–483.
- Hallinan, M. T., & Dame, N. (1996). Track mobility in secondary school. *Social Forces, 74*, 983–1002.
- Helmke, A., Hosenfeld, I., & Schrader, F.-W. (2004). Vergleichsarbeiten als Werkzeug für die Verbesserung der diagnostischen Kompetenz von Lehrkräften [Standardized exams as tools to improve teachers' diagnostic competence]. In R. Arnold & C. Griese (Eds.), *Schulleitung und Schulentwicklung* (pp. 119–144). Hohengehren: Schneider.
- Herppich, S., Praetorius, A.-K., Hetmanek, A., Glogger-Frey, I., Ufer, S., Leutner, D., et al. (2017). Ein Arbeitsmodell für die empirische Erforschung der diagnostischen Kompetenz von Lehrkräften [A working model for empirical research concerning the diagnostic competence of teachers]. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* (pp. 75–94). Münster: Waxmann.
- Hofer, M. (1981). Die Schülerspezifizität in Einstellungen und Verhaltensweisen des Lehrers [Student specificity in teachers' attitudes and behavior]. In M. Haidl (Ed.), *Lehrerpersönlichkeit und Lehrerrolle im sozial-integrativen Unterricht* (pp. 58–82). München: Lurz.
- Hoffrage, U., & Reimer, T. (2004). Models of bounded rationality: The approach of fast and frugal heuristics. *Management Review, 15*, 437–459.
- Hörstermann, T., & Krolak-Schwerdt, S. (2012). Teachers' typology of student categories. A cluster analytic study. In W. Gaul, A. Geyer-Schulz, L. Schmidt-Thieme, & J. Kunze (Eds.), *Studies in classification, data analysis and knowledge organization. Challenges at the Interface of data analysis, computer science, and optimization* (pp. 547–556). Berlin: Springer.
- Hörstermann, T., Pit-ten Cate, I. M., Krolak-Schwerdt, S., & Glock, S. (2017). Primacy effects in attention, recall and judgment patterns of simultaneously presented student information: Evidence from an eye-tracking study. In L. R. Vogel & C. Jenkins (Eds.), *Student achievement: Perspectives, assessment and improvement strategies* (pp. 1–28). Hauppauge: Nova Science.
- Ingenkamp, K., & Lissmann, U. (Eds.). (2008). *Lehrbuch der Paedagogische Diagnostik* [Textbook of educational assessment] (6th ed.). Weinheim: Beltz.
- Jeynes, W. H. (2005). A meta-analysis of the relation of parental involvement to urban elementary school student academic achievement. *Urban Education, 40*, 237–269.
- Jussim, L. (2005). Accuracy in social perception: Criticisms, controversies, criteria, components, and cognitive processes. *Advances in Experimental Social Psychology, 37*, 1–93.
- Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*, 131–155.
- Jussim, L., Eccles, J., & Madon, S. J. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology, 28*, 281–388.

- Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction*, 28, 73–84.
- Kaufman, J. E., & Rosenbaum, J. E. (1992). The education and employment of low-income black youth in white suburbs. *Educational Evaluation and Policy Analysis*, 14, 229–240.
- Klapproth, F., Glock, S., Böhmer, M., Krolak-Schwerdt, S., & Martin, R. (2012). School placement decisions in Luxembourg: Do teachers meet the education Ministry's standards? *The Literacy Information and Computer Education Journal*, 1, 765–771.
- Klapproth, F., Glock, S., Krolak-Schwerdt, S., Martin, R., & Böhmer, M. (2013). Prädiktoren der Sekundarschulempfehlung in Luxemburg: Ergebnisse einer Large Scale Untersuchung [Predictors of recommendations for secondary school type in Luxembourg: Results of a large scale study]. *Zeitschrift für Erziehungswissenschaft*, 16, 355–379.
- Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2013). The impact of accountability on teachers' assessments of student performance: A social cognitive analysis. *Social Psychology of Education*, 16, 215–239.
- Kulik, C. L. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal*, 19, 415–428.
- Lee, Y.-T., McCauley, C., & Jussim, L. (2013). Stereotypes as valid categories of knowledge and human perceptions of group differences. *Social & Personality Psychology Compass*, 7, 470–486.
- Lerner, J. S., & Tetlock, P. E. (1999). Accounting for effects of accountability. *Psychological Bulletin*, 125, 255–275.
- Lucas, S. R. (1999). *Tracking inequality: Stratification and mobility in America's high schools*. New York: Teachers College Press.
- Lucas, S. R. (2001). Effectively maintained inequality: Education transitions, track mobility, and social background effects. *American Journal of Sociology*, 106, 1642–1690.
- Macrae, C. N., Stangor, C., & Hewstone, M. (Eds.). (1996). *Stereotypes and stereotyping*. New York: Guilford Press.
- Martin, R., Dierendock, C., Meyers, C., & Noesen, M. (2008). *La place de l'école dans la société luxembourgeoise de demain* [The position of the school in the Luxembourgish society of tomorrow]. Brussels: De Boeck.
- Martin, R., Ugen, S., & Fischbach, A. (2015). *Épreuves Standardisées: Bildungsmonitoring für Luxemburg. Nationaler Bericht 2011 bis 2013* [Standardised achievement test: Educational monitoring for Luxembourg]. Esch/Alzette: University of Luxembourg.
- McCombs, R. C., & Gay, J. (1988). Effects of race, class and IQ information on judgments of parochial grade school teachers. *The Journal of Social Psychology*, 128, 647–652.
- Nölle, I., Hörstermann, T., Krolak-Schwerdt, S., & Gräsel, C. (2009). Relevante diagnostische Informationen bei der Übergangsempfehlung – die Perspektive der Lehrkräfte [Relevant diagnostic information for school placement decisions – teachers' perspectives]. *Unterrichtswissenschaft*, 37, 294–310.
- Oakes, J. (2005). *Keeping track: How schools structure inequality* (2nd ed.). New Haven: Yale University Press.
- Oakes, J., & Guiton, G. (1995). Matchmaking: The dynamics of high school tracking decisions. *American Educational Research Journal*, 32, 3–33.
- OECD. (2010). *PISA 2009 results: Overcoming social background – Equity in learning opportunities and outcomes (Volume II)*. Paris: OECD.
- Parks, F. R., & Kennedy, J. H. (2007). The impact of race, physical attractiveness, and gender on education majors' and teachers' perceptions of student competence. *Journal of Black Studies*, 37, 936–943.
- Pendry, L. F., & Macrae, C. N. (1996). What the disinterested perceiver overlooks: Goal-directed social categorization. *Personality and Social Psychology Bulletin*, 22, 249–256.

- Pietsch, M., & Stubbe, T. C. (2007). Inequality in the transition from primary to secondary school: School choices and educational disparities in Germany. *European Educational Research Journal*, 6, 424–445.
- Pit-ten Cate, I. M., & Hörstermann, T. (2012). Towards a criterion to judge the accuracy of transition decisions. In *The need for educational research to champion freedom, education and development for all* (p. 66). Cadiz: EERA.
- Pit-ten Cate, I., & Krolak-Schwerdt, S. (2016, September). Übergang in die Sekundarschule: Die Rolle der Entscheidungsverantwortung im Orientierungsprozess [Transition into secondary school: The role of accountability in the tracking process]. *FORUM Für Politik, Gesellschaft Und Kultur* (365), 8–11.
- Pit-ten Cate, I. M., Krolak-Schwerdt, S., Hörstermann, T., & Glock, S. (2013). Better decisions through science – Changing decision making processes by applying formal decision rules. In Ohle, A., & McElvany (Eds.), *Teachers' competencies and teacher judgments*. Symposium conducted at the 15th Biennial EARLI Conference for Research on Learning and Instruction, Munich, Germany.
- Pit-ten Cate, I. M., Krolak-Schwerdt, S., Glock, S., & Markova, M. (2014). Improving teachers' judgments. Obtaining change through cognitive processes. In S. Krolak-schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 45–61). Rotterdam: Sense.
- Pit-ten Cate, I. M., Krolak-Schwerdt, S., Hörstermann, T., & Glock, S. (2015). Assessing teachers' diagnostic competence: Predictive validity and application of a criterion to judge the accuracy of transition decisions. In Pant, H. A. & Zlatkin-Troitschanskaia, O (Eds.), *Modeling and measuring academic competencies in higher education*. Symposium conducted at the European Conference on Educational Research, Budapest, Hungary.
- Pit-ten Cate, I. M., Krolak-Schwerdt, S., & Glock, S. (2016a). Accuracy of teachers' tracking decisions: Short- and long-term effects of accountability. *European Journal of Psychology of Education*, 31, 225–243.
- Pit-ten Cate, I. M., Krolak-Schwerdt, S., Hörstermann, T., & Glock, S. (2016b). *Theoretical knowledge and formal decision rules: Can we reduce bias in orientation decisions?* Paper presented at the 4. Tagung der Gesellschaft für Empirische Bildungsforschung, Berlin.
- Rumberger, R. W., Ghatak, R., Poulos, G., Ritter, P. L., & Dornbusch, S. M. (1990). Family influences on dropout behavior in one California high school. *Sociology of Education*, 63, 283.
- Schalke, D., Brunner, M., Geiser, C., Preckel, F., Keller, U., Spengler, M., & Martin, R. (2013). Stability and change in intelligence from age 12 to age 52: Results from the Luxembourg MAGRIP study. *Developmental Psychology*, 49, 1529–1543.
- Seidel, T., & Prenzel, M. (2007). Wie Lehrpersonen Unterricht wahrnehmen und einschätzen – Erfassung pädagogisch-psychologischer Kompetenzen mit Videosequenzen [*How teachers perceive lessons – Assessing educational competencies by means of videos*]. *Zeitschrift für Erziehungswissenschaft*, 8, 201–216.
- Seidel, T., Stürmer, K., Blomberg, G., Kobarg, M., & Schwindt, K. (2011). Teacher learning from analysis of videotaped classroom situations: Does it make a difference whether teachers observe their own teaching or that of others? *Teaching and Teacher Education*, 27, 259–267.
- Stangor, C., & Schaller, M. (1996). Stereotypes as individual and collective representations. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 3–40). New York: Guilford Press.
- Stubbe, T. C., & Bos, W. (2008). Schullaufbahnenempfehlungen von Lehrkräften und Schullaufbahnentscheidungen von Eltern am Ende der vierten Jahrgangsstufe [Teacher's school track recommendations and parents' school track decisions at the end of primary school]. *Empirische Pädagogik*, 22, 49–63.
- Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum [The simulated classroom]. *Zeitschrift Für Pädagogische Psychologie*, 22, 261–276.
- Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, 104, 743–762.

- Swets, J. A., Dawes, R. M., & Monahan, J. (2000, October). Better decisions through science. *Scientific American*, 283, 82–87.
- Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 45, 74–83.
- Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 331–376). New York: Academic.
- Tetlock, P. E., & Lerner, J. S. (1999). The social contingency model: Identifying empirical and normative boundary conditions on the error-and-bias portrait of human nature. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology*. New York: Guilford Press.
- Thill, M. (2001). *La nouvelle procédure de passage de l'enseignement primaire (public et privé) vers l'enseignement secondaire et secondaire technique : Resultats des conseil d'orientation et des procédures de recours* [The new transition from primary to secondary school : Results of orientation committee and appeal procedures]. Luxembourg: MENFP, SCRIPT.
- Van de Werfhorst, H. G., & Mijs, J. J. B. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology*, 36, 407–428.
- van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, 47, 497–527.
- Weiss, H. B. (1989). State family support and education programs: Lessons from pioneers. *American Journal of Orthopsychiatry*, 59, 32–48.

Chapter 6

Threshold Concepts for Modeling and Assessing Higher Education Students' Understanding and Learning in Economics



Sebastian Brückner and Olga Zlatkin-Troitschanskaia

Abstract In the last decade, the research carried out on threshold concepts as a content-based way to model students' understanding and learning in several domains has increased. However, empirical evidence on this approach is still scarce. In this chapter, the authors investigate the adequacy of the threshold concepts approach in the domain of business and economics in higher education following an established differentiation between basic, discipline, and modeling thresholds. After conducting a cognitive interview study using verbal reports, a self-assessment questionnaire was used to assess the respondents' familiarity with the content and their security to solve the tasks. Results indicate that there is a complex relation between students' response processes, self-assessment, and test scores, which varies according to the different thresholds and that all three measures generally confirm our hypotheses yet have to be critically discussed. There are implications that test developers, test users, respondents, and other stakeholders should be aware of this complex relation; it affirms that the threshold concepts approach is at least a useful tool when conceptualizing and developing tests, which can be considered to be an addition to classic taxonomies of educational objectives.

6.1 Introduction and Objectives

The assessment of academically acquired knowledge is becoming more and more important in higher education. Standardized testing instruments used in this context often follow a cognitive modeling approach, in which knowledge acquisition is mostly explained with subject-specific cognitive dispositions that can be depicted as a continuum of gradual and content characteristics (see, a model by Zlatkin-Troitschanskaia et al. 2014). While data on content structures is available and often

S. Brückner (✉) · O. Zlatkin-Troitschanskaia
Johannes Gutenberg University, Mainz, Germany
e-mail: brueckner@uni-mainz.de; lstroitschanskaia@uni-mainz.de

seems to follow a rather functional, content-oriented, or curricular outline, research on the differentiation between the various cognitive levels of knowledge is still lacking (e.g., Walstad 2001). In order to model the gradual expression of academically acquired knowledge, a criterial assessment reference standard is necessary, for example, to document study progress or employability and to pass this data on to students, universities, institutions, etc. (Allgood and Bayer 2016; Macha and Schuhen 2011; Zumbo 2016). This can be achieved by employing gradual modeling approaches in research, which reference classic taxonomies of cognitive processes (e.g., reproduction, application, evaluation) or types of knowledge (e.g., strategic knowledge, procedural knowledge, conceptual knowledge) (e.g., Bloom et al. 1956; Anderson and Krathwohl 2001; Biggs and Collis 1982). However, these taxonomies often do not correspond with the expectations of their expressions as formulated prior to test development (Schumann and Eberle 2011; Gierl 1997).

While traditional theories of cognitive processes differentiate between understanding and applying knowledge, thus referring back to Bloom (Anderson and Krathwohl 2001), alternative learning theories emphasize that every learning process requires subject-specific knowledge of the domain-specific concepts (e.g., principles, facts, rules) (Davies 2012, p. 251) or focus more on the inferential construction of domain-specific learning (e.g., Biggs and Collis 1982; Minnameier 2013; see also Alexander, Chap. 3 in this volume). One example is the “threshold concepts” approach, which describes learning progress by focusing on fundamental concepts that determine knowledge and understanding in a specific academic domain (Davies 2012).

In order to explore the characteristics of such an approach for modeling understanding and learning in a specific domain, it is necessary to carefully analyze mental operations that occur along such thresholds. Besides the primarily cognitive operations (such as remembering, applying, analyzing), which are already included in the classic taxonomies, studies have so far not been able to show that learning success, which is expressed in examination results, is not only influenced by cognitive operations (e.g., Liu et al. 2012; Musekamp and Pearce 2016). The level of noncognitive states (e.g., the perceived familiarity that participants assign to the respective contents) is highly relevant for solving domain-specific problems and tasks (cf. Alexander, Chap. 3 in this volume). Moreover, test results and the assessment of knowledge and understanding can be confounded by, for example, heuristic decision-making patterns (e.g., Krolak-Schwerdt et al., Chap. 5 in this volume) or the application of test-taking strategies (e.g., random guessing). It is possible, for example, that the more complex a decision on a final response to a task is perceived to be, the higher the instances of guessing strategies are. Therefore, it is highly important to closely analyze these mental operations with regard to their correlation with the specified threshold concepts.

In the following, this challenge shall be addressed using the so-called “threshold concepts” approach. To avoid the deficits of modeling approaches based on traditional taxonomies as described above, the theory of threshold concepts by Meyer and Land (2005) (Davies 2012, p. 251) is beneficial as it depicts knowledge and understanding directly based on the contents and concepts of a domain. The approach is examined in a microanalytic study with regard to its potential suitability for mod-

eling understanding and learning, here, in the domain of business and economics, using objectively determined mental operations as well as students' attitudes toward the response process as described in self-assessment questionnaires. Particularly in the domain of business and economics, the testing instruments used in higher education show that the prediction of test scores at different levels, on the basis of such taxonomies, is not always successful. Discussions as to why that may be revolve around the issues of the partially difficult measurability of cognitive processes and types of knowledge as well as overly abstract cross-discipline-oriented taxonomies. Several studies on economic knowledge in higher education (Asano and Yamaoka 2015; Jang et al. 2010; Walstad et al. 2007; Hahn et al., Chap. 8 in this volume) have attempted to classify the scores according to these taxonomies. Particularly, the Test of Understanding in College Economics (TUCE) (Walstad et al. 2007) and the Examen General para el Egreso de Licenciatura en Administración (EGEL-A) as well as the Examen General para el Egreso de Licenciatura en Contabilidad (EGEL-C) (Uribe 2013), two tests that are frequently used in higher education assessment, are lacking a cross-national comparative approach for criterion-based modeling of understanding and learning in business and economics that goes beyond rough modeling approaches, for example, the taxonomy of Bloom et al. (1956).

After drafting the threshold concepts approach as well as its application to understanding and learning in business and economics, current studies that empirically examine this approach will be presented. Afterwards, links to mental operations related to the thresholds will be put forward. Subsequently, four hypotheses that provide first insights regarding the approach's suitability when it comes to gradually modeling understanding and learning will be formulated (see Sect. 6.2). In order to test these hypotheses and gain an insight into the mental operations, cognitive interviews using the think-aloud method were conducted with 20 students, while they were working on 19 tasks in a business and economics test (Brückner 2017). In addition, a self-assessment questionnaire was used to collect data on how students dealt with each task (see Sect. 6.3). In order to confirm these hypotheses, the empirical results, the link between threshold concepts and final responses, mental operation data, and self-assessment will be presented in detail (Sect. 6.4). Finally, the study's results as well as their implications for common approaches for modeling levels and grades of business and economic knowledge and understanding will be critically discussed (Sect. 6.5).

6.2 State of Research in Higher Education

6.2.1 *Threshold Concepts for Modeling Knowledge, Understanding, and Learning*

The idea of threshold concepts stems from the research project ETL (Enhancing Teaching and Learning Environments in Undergraduate Courses Project), which focuses on identifying factors for highly qualified learning environments (Meyer

and Land 2003, p. 1). In 2000, Meyer introduced the so-called threshold concepts in order to differentiate between learning contents that enabled the understanding of issues and concepts and those that did not result in any change of conception (O'Donnell 2009, p. 191). Meyer and Land's (2005) theory of threshold concepts helps characterize learners' understanding and learning processes within a certain discipline (Davies and Mangan 2007, p. 3). They focus on describing specific concepts of a discipline or content area and incorporate both the social and cognitive dimensions of learning (Davies 2012, p. 250). In this sense, it is assumed that, once they have been understood, these concepts will gradually enable access to thoughts that were otherwise inaccessible (Kricks et al. 2013, p. 18). With regard to the learning process, threshold concepts represent an essential transformation in the way of thinking, interpreting, and perceiving (Meyer and Land 2006, p. 3). What is fundamental about the theory of threshold concepts is the notion that the most crucial ideas of a discipline cannot be simplified and are thus inaccessible to novices (Davies 2012, p. 250). By revising both existing day-to-day and previously acquired knowledge, learners start to develop a transitional understanding that incorporates both specialist knowledge and explanations for discipline-specific concepts (Davies 2012, p. 250; Davies and Mangan 2007, p. 3). This means that learners of economics have to, for example, develop an understanding of the fact that technical terms such as "cost" and "invest" mean something else than when used in their everyday sense (Davies 2012, p. 250).

In this respect, Meyer and Land (2006) identified five characteristics of threshold concepts: *transformative*, *irreversible*, *integrative*, *bounded*, and *troublesome*. Threshold concepts are *transformative* because they significantly change the way learners understand, interpret, and perceive content – an aspect crucial to their further progression in their learning process. This characteristic reflects the interface function of threshold concepts, as learners must first exceed these to be able to penetrate a subject matter accordingly (Meyer and Land 2006, pp. 7–8). Threshold concepts are *irreversible* due to the fact that the fundamental understanding of these concepts cannot be reversed and functions as a basis to understanding other crucial concepts (Meyer and Land 2006, p. 7). The aspect of irreversibility is particularly significant for teaching, as both teachers and learners often exhibit difficulties returning to a certain threshold they crossed a long time ago. Threshold concepts are *integrative*, as they outline previously hidden links and correlations to the learner. Thus, threshold concepts can reveal common contents and conceptual links between subdomains of one discipline that were previously regarded as completely disconnected (Meyer and Land 2006, p. 7). One function of threshold concepts is that they open up new learning spaces, thus *bounding* certain content areas, which may help to connect or disconnect disciplines (Meyer and Land 2006, p. 8). Because of their high significance for learning within a certain discipline, threshold concepts present a relatively high number of challenges to the learners and can be quite *troublesome* to understand, as students are required to overcome the boundaries of their understanding (Meyer and Land 2006, p. 8; O'Donnell 2009, p. 191).

The term threshold concept must be differentiated from other types of concepts (Meyer and Land 2003) such as core concepts or key concepts (Davies 2012).

A core concept or a key concept is understood as a conceptual basic element, which develops a learner's understanding of a topic (Meyer and Land 2003, p. 4). This basic element must be understood – however, understanding it does not lead to a differentiated perception of other concepts or content previously learned (Meyer and Land 2003, p. 4). While core and key concepts perceive the learning process as an accumulation of knowledge and can be instructionally simplified, threshold concepts encompass a structural transition (Davies 2012). Threshold concepts are particularly relevant for learning as they are said to enable the development of a fundamental understanding of a domain. This, alongside a change of perspective, is often referred to as “conceptual change” (Davies and Mangan 2007, p. 3; Kricks et al. 2013, p. 21) or “conceptual learning” (Gagné 1985) and is characterized by the differentiation of different types of thresholds (Davies and Mangan 2007).

One possible approach to hierarchizing these sources can be found in Davies and Mangan (2007, p. 4) as well as, referring to their work, in Kricks et al. (2013, p. 21). The authors distinguish the following types of conceptual change, which are passed sequentially during the learning process: *basic*, *discipline*, and *modeling*. Learners passing the threshold of basic concepts are able to assign a new, domain-specific meaning according to the basic principles and explanations of a discipline to concepts that they could previously only understand based on their day-to-day experiences. The concepts along this (lower) *basic threshold* are often much frequented through everyday knowledge. However, without the appropriate understanding, they are often misinterpreted in the domain context; thus, threshold concepts present a first, fundamental level of access to a domain's content (Davies 2012). Subsequently, *discipline concepts* enable learners to also understand concepts from a theoretical, domain-specific perspective. The majority of these concepts are not familiar from day-to-day life, and thus a false or naïve understanding of them frequently exists; the understanding of these concepts has to be developed through instructional measures. The third type of conceptual change, the *modeling concept*, enables the learner to understand specific rather abstract modeling approaches within a domain and to apply them in argumentative discourse in order to further develop theories within the discipline. This requires a scientific approach to the relevant theories and models of a domain so that they can be developed, assessed, and reviewed (Davies and Mangan 2007, p. 4). Once the threshold of modeling concepts is crossed, the learners have access to an elaborate understanding of a domain.¹

Despite the fact that this approach has been tested in various disciplines (e.g., mathematics and education science), there have been hardly any empirical investigations into the theory of threshold concepts (Riegler 2014; Shanahan et al. 2006). First conceptual approximations regarding the threshold concept of opportunity costs already exist in economics (Davies and Mangan 2007; Davies 2012; Shanahan et al. 2006); however, there are only very few empirical studies on threshold concepts in business and economics in higher education (see Table 6.1).

¹For a comparison between the gradual development of knowledge according to the taxonomy of Bloom et al. (1956) and the approach of threshold models, see also Davies (2012).

Table 6.1 Empirical studies on threshold concepts in business and economics in higher education

Author	Research focus	Sample	Result
Kricks et al. (2013)	Threshold concepts in economics (e.g., opportunity cost)	16 students in economics	Due to the heterogeneity of the meanings mapped to the concepts, this approach only makes sense if applied in combination with a taxonomy of teaching-learning objectives (SOLO) (Biggs and Collis 1982)
Lucas and Mladenovic (2009)	Threshold concepts in business (e.g., cash and profit, depreciation)	98 first- and second-year students, England and Australia	The students' economic concept knowledge does not meet the expectations specified in the respective curricula
Davies and Mangan (2007)	Threshold concepts in economics (e.g., opportunity cost, comparative advantage)	12 university lecturers and more than 20 economic students	Lecturers as experts use far more concepts when explaining economics principles than students, while students mainly follow rather simple one-dimensional lines of argumentation
Reimann and Jackson (2006)	Threshold concepts in economics (e.g., opportunity cost and elasticity)	30 first-year students (2001–2003)	Everyday phenomena can be used to assess the lower thresholds (e.g., basic and discipline) in economic science
Shanahan et al. (2006)	Threshold concepts in economics (e.g., opportunity cost)	700 students, 40 multiple-choice questions in microeconomics course	Weak relationship between course performance and understanding of the threshold concept
Meyer and Land (2006)	Threshold concepts in economics (e.g., opportunity cost)	30 students from an introduction to microeconomics course	Highlights the importance of threshold concepts for the design of teaching-learning environments and provides suggestions on how to effectively embed these into teaching

Moreover, these few existing pieces of evidence are rather inconsistent. While some studies show proof of an expected shift in the sequence of conceptual change from basic to modeling (Davies and Mangan 2007, 2009; Meyer and Land 2006; Reimann and Jackson 2006), other studies expand the existing modeling, based on conceptual change, with taxonomic categories such as the SOLO (Structure of the Observed Learning Outcome) taxonomy (Kricks et al. (2013); Lucas and Mladenovic 2009).²

The German Kricks et al. (2013) study examines the threshold concepts approach in economics with a subsample of 16 economics students. The authors conclude that due to the heterogeneity of the meanings mapped to the concepts, this approach only makes sense if applied in combination with a taxonomy of teaching-learning objectives. Lucas and Mladenovic (2009) applied the threshold concept approach in

² Due to the novelty of this approach, it might be comprehensible that only few threshold concepts have so far been identified and empirically analyzed (Davies 2012).

their study with 98 students in their first and second year of study in England and Australia and concluded that the students' economic concept knowledge does not meet the expectations specified in the respective curricula. In the study conducted by Davies and Mangan (2007), the aim was to investigate the understanding of threshold concepts in economics with a sample of 12 university lecturers and more than 20 students. They found that lecturers as experts use far more concepts when explaining facts than students, while students mainly follow rather simple one-dimensional lines of argumentation. Compared to students, lecturers more frequently use diagrams to explain concepts, while fewer elaborated explanations correlate to use of fewer diagrams. In general, they found that, overall, the lecturers had a more elaborated knowledge base with and about threshold concepts than students. In another study from 2009, the authors interviewed students and presented the learning of the relationships in the IS/LM (investment-saving/liquidity preference-money supply) model as well as the learning of overall supply and demand as essential for acquiring knowledge in the economic science study program (Davies and Mangan 2009). In an investigation of 30 first-year students between 2001 and 2003, it was found that everyday phenomena can be used to empirically assess the lower thresholds in economic science (Reimann and Jackson 2006). So far, the most extensive study with about 700 students focused on the specific threshold concept of opportunity costs. The students were given 40 multiple-choice questions in a course on microeconomics. The results suggest a weak relationship between course performance and understanding the threshold concept (Shanahan et al. 2006). In addition, Meyer and Land (2006) conducted a study with 30 students from an introduction course to microeconomics. In their study, they highlight the importance of threshold concepts for the design of teaching-learning environments and provide suggestions on how to effectively embed these into teaching.

In total, quite a few studies have already been conducted on threshold concepts. Overall, there is still a need for further studies that both provide evidence for the suitability of the threshold concept for an appropriate modeling of students' knowledge, understanding, and learning in business and economics and investigate the potential of threshold concepts within teaching and learning in higher education.

6.2.2 Hypotheses

If we apply the hierarchy presented above to the discipline of business and economics, we should be able to see a transition in learners who have already passed and now possess an understanding of the several thresholds. In regard to *basic concepts*, students' everyday and naïve understanding should progress into a rudimentary understanding of economic sciences. For example, learners should be capable of distinguishing between the concepts of price and costs or income and wealth (Davies and Mangan 2007, p. 4). In contrast, learners who have developed an understanding of *discipline concepts* should demonstrate a transition from a rudimentary,

economic understanding to an elaborated knowledge of economic principles and methods. Therefore, they should have an understanding of the interaction of markets and comparative cost advantages (Davies and Mangan 2007, p. 4). The transition to the so-called modeling concepts is characterized by the development of an understanding of contents on all levels, from basic to advanced, discipline-specific expertise. This should result in, for example, an understanding of comparative statics and differing time analyses (short term, medium term, long term) (Davies and Mangan 2007, p. 4). The high expectations attached to conceptual change on higher levels become apparent in the fact that students usually have systematic difficulty in developing an elaborated knowledge base at higher levels and might also need several years of practical experience to generate this expertise. This results in the following hypothesis on the connection between economic knowledge and a gradual structuring of threshold concepts:

H1: Students know less about concepts along higher thresholds (e.g., modeling concepts) than about concepts along lower thresholds (e.g., basic concepts).

Threshold concepts can be considered “neuralgic points” in the development of economic knowledge and, isolate naïve, everyday experiences from specialist knowledge structures. Particularly concepts on lower thresholds that provide students with their first access to domain understanding exhibit this isolation function in regard to everyday experiences (Lucas and Mladenovic 2006, p. 148; Kricks et al. 2013, p. 21). Students’ familiarity with concepts that are closely linked to everyday life (e.g., basic concepts) should be greater than their familiarity with other concepts along higher thresholds. The same also applies to dealing with these concepts.

Although it often requires a conducive situation to judge how confident students are in using the concepts, the higher the familiarity, the higher their confidence tends to be. Simplifying economic principles in their practical use is a well-known technique when solving economic problems and has frequently been linked to the concept of economic heuristics (Brückner and Pellegrino 2016; Leiser and Aroch 2009); it allows for the following assumption:

H2: Students are more familiar with concepts along lower thresholds (e.g., basis and discipline) and are more confident using them for solving tasks than they are with concepts along higher thresholds (e.g., discipline and modeling).

The mental operations that can occur along the thresholds and do not encompass heuristics have also been subject to little research. An important aspect in dealing with various threshold concepts is assigning a meaning to a concept, also in terms of an abductive inference (Minnameier 2013). A common strategy used in the case of not knowing something is “random guessing” (Brückner 2017). It can be assumed that students guess less frequently when it comes to concepts along lower thresholds than concepts along higher ones. At the same time, guessing a meaning along higher thresholds based on the specificity of concepts should not result in as an elaborate understanding as it does along lower thresholds. Based on findings in the field of economic knowledge with standardized tests, it can be assumed (Zlatkin-Troitschanskaia et al. 2014; Walstad et al. 2007) that students are less successful in

solving tasks that include concepts along higher thresholds. This leads to the following assumptions:

H3: Students tend to guess more frequently when solving tasks including concepts of higher thresholds.

H4: Along higher thresholds, the correlation between how often students guess and how often they choose the right solution for an economic task decreases.

6.3 Study Design

6.3.1 Modeling and Measuring of Economics Knowledge

To verify the hypotheses, 19 tasks were chosen from the WiwiKom study³ (Zlatkin-Troitschanskaia et al. 2014). The tasks, which were translated and adapted into German, are part of the internationally established Test of Understanding in College Economics (TUCE) (Walstad et al. 2007) and the Examen General para el Egreso de Licenciatura en Administración (EGEL-A) as well as the Examen General para el Egreso de Licenciatura en Contabilidad (EGEL-C) (Uribe 2013). Each task contains a central business or economic concept that the participants need to understand in order to choose the correct answer from the four options. The concepts were all mapped to a certain threshold so each task corresponds to a threshold. The medium threshold “discipline concepts” is attributed to a high level of ability to differentiate economic knowledge. It is also the first threshold suitable to access subject-specific understanding. Most tasks refer to the medium threshold, which results in the following distribution (see Table 6.2)⁴.

Table 6.3 presents a task for the concept of the “product life cycle” which can be accessed via nominal decomposition but not using aspects of subject-specific correctness. Solving this task requires a discipline-specific understanding of the phases of the product life cycle which can be developed through (institutionalized) learning processes. This includes subject-specific instructions to activate the learner’s learning processes linked to these concepts

These tasks were solved by a subsample of 20 students from the WiwiKom study (Brückner 2017). The test takers were students who on average were in their second year of study (mode = 1; median = 2). Therefore, this study includes students from both the beginning and the end of their studies. The degree of economic knowledge is determined by the number of correctly solved tasks. A correct answer is coded

³Modeling and Measuring Competencies in Business and Economics in Higher Education funded by the German Federal Ministry of Education and Research (BMBF), Grant No: 01PK11013.

⁴Further indices emphasizing the concepts’ relevance and use in day-to-day life could be delivered by an analysis of the frequency with which the concepts are used in web-based search engines (e.g., Google or Yahoo). This analysis shows that 5 million mentions can be found for the aforementioned concepts at the basic concepts threshold, several hundreds of thousands for the discipline concepts, and less than 1 hundred thousand for the modeling concepts.

with a 1, and an incorrect answer is coded with a 0. As the tasks are mapped along different thresholds, the degree of economic knowledge can be mapped to individual thresholds.

Table 6.2 Task distribution along thresholds according to Davies and Mangan (2007)

	Threshold “basic concepts”	Threshold “discipline concepts”	Threshold “modeling concepts”
Number of items (%)	3 (16)	11 (58)	5 (26)
Threshold concepts (Davies and Mangan 2007; Lucas and Mladenovic 2009)	For example, Cash and profit	For example, Depreciation	For example, Comparative statics Intertemporality
	Income	Opportunity cost	
	Investment and saving	Partial equilibrium	
Concepts along the thresholds	For example, Gross salary	For example, Critical path method	For example, Dynamic procedures of investment analysis Strategic human resource planning
	Optimization	Growth-share matrix	
	Revenue	SWOT matrix	
		Diamond model (Porter)	
		Product life cycle	
		Leverage effect	
Break-even point			

Table 6.3 Threshold “discipline concepts”

A company has reached a 20% increase in sales for one of its products. Due to increasing competition on the market, the company decides to use 5% of its profit to reinforce their advertising efforts.

Please name the stage of the product life cycle that this process refers to.

Maturity

Growth

Decline

Introduction

6.3.2 Familiarity and Confidence

After completing each task, the students were asked how familiar they were with the concepts and how confident they were in solving the tasks. To this end, two factors were noted after each task in the form of a six-level Likert scale (not familiar = 1 up to very familiar = 6 and not confident = 1 up to very confident = 6). The questions

were: “How familiar are you with the content of the question?” and then “How confident were you answering the question?” The aim was to identify deviations between an understanding of the concepts in the task on the one hand, and, on the other hand, further aspects that play a significant role in solving the task and can therefore be predominantly associated with confidence in problem-solving. As the items were all constructed in a closed-ended format with one question and four response options with a dominant content-related concept (e.g., the Break-even point or the aggregated demand), familiarity was only perceived with the content of the task and not, for example, the item or test format.

6.3.3 Guessing Behavior

During the process of working on the tasks, the students’ guessing behavior was also assessed. A large number of methods exist for this purpose, ranging from simple surveys using interview techniques to complex neurobiological methods. In the study in question, students were asked to verbalize their thoughts using the concurrent think-aloud method while solving the tasks (Ericsson and Simon 1993). This was intended to assist in the manifestation of hints on mental operations that provide information on guessing behavior, so that these could then be depicted in a numeric relative. While solving the tasks, students were not interrupted and were only spoken to when being asked to think out loud by saying “please continue to speak” (Brückner 2017).

6.4 Results

In total, there were 380 individual task solutions, 60 of which are mapped along the threshold “basic concepts” ($n = 60$), 220 along the threshold “discipline concepts” ($n = 220$), and 100 along the threshold “modeling concepts” ($n = 100$) (see Table 6.4).

Table 6.4 Proportions of correct final responses along the thresholds

Thresholds	WiwiKom substudy				WiwiKom study t1				
	<i>M</i>	<i>p/n</i>	SE	[95% CI]		<i>M</i>	SE	[95% CI]	
Basic concepts	0.783	47/60	0.054	0.678	0.889	0.673	0.021	0.631	0.715
Discipline concepts	0.505	111/220	0.034	0.438	0.571	0.484	0.007	0.469	0.498
Modeling concepts	0.460	46/100	0.050	0.362	0.559	0.394	0.011	0.373	0.416

Note: *p* is the number of correct responses on each threshold. *n* is the total number of responses on each threshold

The average solution frequency of the individual thresholds offers first insights into the question as to whether students really have a higher level of knowledge about concepts along lower thresholds. It becomes evident that students apparently have a higher level of knowledge along the threshold “basic concepts,” with a frequency of correct final responses of 78.33%, than along the other two thresholds, for which a frequency of correct final responses of 50.45% (discipline concepts) and 46% (modeling concepts) was determined. This corresponds to 47 correct final responses on the threshold “basic concepts,” 111 on the threshold “discipline concepts,” and 46 on the threshold “modeling concepts.” The confidence intervals offer additional evidence of a highly significant ($p < 0.01$) difference between the lowest and the other two thresholds but not between the two thresholds “discipline concepts” and “modeling concepts,” as these intervals overlap (see also Cumming and Finch 2005).

The additional analysis of the proportions of correct responses of the selected tasks from the overall study t1 ($N = 3783$) provides indications for how the findings can be generalized. Due to the interview situation and individual processing, the proportions correct in the substudy are more pronounced than in the overall study. However, it is evident that the difference between the thresholds is even more obvious and that there are significant differences between the three thresholds (Table 6.4).

The analysis of statements on familiarity with the concepts indicates that students are more familiar with concepts along the threshold “basic concepts” ($M = 3.96$) than with concepts along the other two thresholds (discipline concepts: $M = 3.30$; modeling concepts: $M = 2.86$) (see Table 6.5 and Fig. 6.1).

Table 6.5 Familiarity and confidence along the thresholds

	Familiarity				Confidence			
	<i>M</i>	SE	95% LL	95% UL	<i>M</i>	SE	95% LL	95% UL
Basic concepts	3.962	0.241	3.488	4.435	4.056	0.223	3.616	4.495
Discipline concepts	3.298	0.119	3.064	3.532	2.884	0.108	2.671	3.097
Modeling concepts	2.856	0.165	2.532	3.179	2.333	0.147	2.045	2.622

The participants are therefore more familiar with content that is conceptually more oriented toward everyday matters and less toward subject-specific matters. In comparison to confidence in the solution,⁵ similar findings regarding the manifestation of the two properties across thresholds become evident. However, the manifestation of confidence in the solution across the thresholds is significantly different, whereas for familiarity with the concepts there is only a significant difference between the thresholds “basic concepts” and “modeling concepts.” Another striking finding in comparing the two characteristics is that on the threshold “basic concepts,” the con-

⁵Familiarity with a concept correlates with the confidence in the solution of strongly concept-dependent tasks. With a correlation $r = 0.73$ between both characteristics, unity of the characteristics cannot necessarily be assumed. As they refer to different phases of a task solving process (familiarity refers to the perceived content and confidence to the final solution of a task), a content-related separation is necessary.

confidence in the solution is slightly higher than the familiarity with the concepts (difference $BC_{s-v} = 0.094$), whereas the opposite is the case for the other two thresholds and the difference increases with higher thresholds (difference $DC_{s-v} = -0.414$; difference $MC_{s-v} = 0.523$) (Table 6.5 and Fig. 6.1). It is obvious that, in addition to familiarity with the concepts, other components of the tasks influence the task solving behavior. Familiarity with the concepts strongly correlates with the confidence in the solution. The fact that students are more familiar with concepts along lower thresholds than with concepts along higher thresholds demonstrates the accessibility and therefore also the importance of threshold concepts for teaching and learning economic contents. Causes for the increasing difference are still unclear, and the mental operations that are related to the threshold concepts must be analyzed more thoroughly. A first step toward achieving this goal will be presented in the following. Students' guessing behavior, which can indicate a lack of economic knowledge and understanding, will also be analyzed in relation to the thresholds and task solutions.

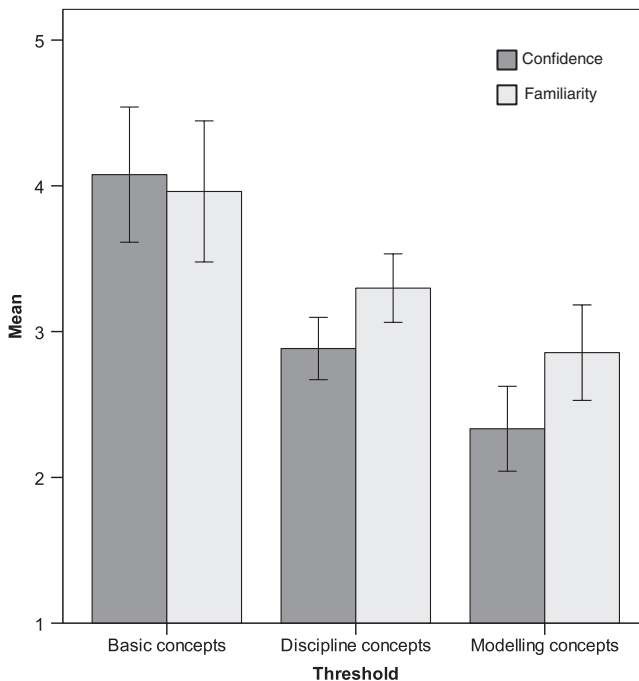


Fig. 6.1 Familiarity and confidence along the thresholds

Guessing behavior was coded dichotomously the same way the final response was coded (0 = not guessed, 1 = guessed), so that the average of all tasks along a threshold maps the guessing probability along this threshold. Across the thresholds, a steady increase of the probability of guessing can be noted. Along the threshold

“basic concepts,” the guessing probability is relatively low at almost 7%, and only 4 out of 60 task solutions are guessed. Along the threshold “discipline concepts,” the guessing probability already increases to 18% (41 guessed solutions in 220 task solving processes), and along the threshold “modeling concepts,” it increases to 33% (33 out of a 100) (Table 6.6). As was already the case for the final responses, it also becomes evident for guessing that the guessing probability along the lowest threshold significantly differs from that of the two higher thresholds; the significance threshold between “discipline concepts” (95% CI UL = 0.238) and “modeling concepts” (95% CI LL = 0.237) is only narrowly missed. The rank correlation indicates a lower but highly significant correlation ($\rho = 0.21$).

Table 6.6 Guessing proportions along the thresholds

Thresholds	Guessing				
	<i>M</i>	<i>g/n</i>	SE	[95% CI]	
Basic concepts	0.067	4/60	0.033	0.003	0.131
Discipline concepts	0.186	41/220	0.026	0.135	0.238
Modeling concepts	0.330	33/100	0.047	0.237	0.423

Note: *g* is the number of guesses along each threshold. *n* is the total number of responses along each threshold

If the students’ final responses are included in addition to the thresholds, a 2×2 matrix can be determined for every threshold, and the relation to the solutions can be analyzed for the individual thresholds. In accordance with expectations, negative correlations with the task solutions should be expected (0 = incorrect solution, 1 = correct solution). The guessing efficacy, defined as the proportion of a successful final response when applying a guessing strategy per threshold, should be lower at higher thresholds than at lower thresholds. For the threshold “basic concepts,” it becomes evident that guessing behavior is associated with a 50% chance of a correct solution and a 50% chance of an incorrect solution. A nonsignificant positive effect ($\omega = 0,184$) is determined (Cohen 1988) which means that guessing is neither a beneficial nor inhibiting factor in the task solving process (Table 6.7).⁶

Table 6.7 Guessing X final response on the threshold “basic concepts”

Guessing			Final response		
			False	Correct	Total
Absent	<i>n</i>		11	45	56
		%	18.33	75.00	93.33
Present	<i>n</i>		2	2	4
		%	3.33	3.33	6.67
Total	<i>n</i>		13	47	60
		%	21.67	78.33	100.00

$\chi^2(df) = 2.021(1); p = 0.155; \omega = 0.184$

⁶As the cell allocation falls short of the required 5% cell frequency, a Fisher-Freeman-Halton test was used as well, which also shows the insignificance of the correlation ($p = 202$) (Lydersen et al. 2007).

For the threshold “discipline concepts,” all cell frequencies exceed the required 5% threshold, and a highly significant negative effect ($\omega = -0,203$) can be determined, indicating that a more frequent application of guessing strategies correlates with more incorrect final responses (Table 6.8). Compared to the “basic concepts,” along the threshold “discipline concepts,” participants guessed more often.

The correlation between guessing and the final response for concepts on the threshold “modeling concepts” is also highly significant and negative ($\omega = -0.349$) and 0.146 points lower than the effect of “discipline concepts.” The strength of the correlation is greater for “modeling concepts,” as 26 of 33 applications of guessing strategies led to an incorrect solution (Table 6.9). Overall, across the thresholds, the correlation between erroneous final responses and applications of guessing strategies increases according to expectations. Dealing with discipline-specific concepts in tasks therefore appears to be a greater challenge for students than dealing with concepts that are more relevant to everyday life.

Table 6.8 Guessing X final response on the threshold “discipline concepts”

Guessing			Final response		
			False	Correct	Total
Absent	<i>n</i>		80	99	179
		%	36.36	45.00	81.36
	Present	<i>n</i>	29	12	41
		%	13.18	5.45	18.64
Total	<i>n</i>	109	111	220	
	%	49.55	50.45	100.00	

$$\chi^2(\text{df}) = 9.048(1); p < 0.010; \omega = -0.203$$

Table 6.9 Guessing X final response on the threshold “modeling concepts”

Guessing			Final response		
			False	Correct	Total
Absent	<i>n</i>		28	39	67
		%	28.00	39.00	67.00
	Present	<i>n</i>	26	7	33
		%	26.00	7.00	33.00
Total	<i>n</i>	54	46	100	
	%	54.00	46.00	100.00	

$$\chi^2(\text{df}) = 12.183(1); p < 0.001; \omega = -0.349$$

6.5 Discussion

This study examined the modeling of threshold concepts in business and economics using the concepts from a selection of standardized items from the WiwiKom study (Zlatkin-Troitschanskaia et al. 2014). In this study, the correlations between tasks corresponding with the different thresholds and guessing behavior, familiarity with the concepts as well as response behavior and confidence in responding were

analyzed. The findings also allow for insights into different mental operations that can be associated with these concepts and outline the workings and potential of this modeling approach for teaching and learning as well as for constructing assessments from different perspectives.

With regard to designing curricula (see also Davies 2012), threshold concepts provide multiple approaches for teachers by outlining “sensitive learning points” and are therefore of central significance for the structure and order of curricular contents. This is indicated by the gradually decreasing response capabilities of the students subjected to this substudy as well as the WiwiKom subsample of 3783 students along the defined thresholds. In accordance with existing findings, the results of this study clearly illustrate that learners’ familiarity with the concepts and thus their previously acquired domain-specific knowledge must be taken into account when planning and implementing instructional measures (see also Alexander, Chap. 3 in this volume). Differentiating between basic, discipline, and modeling thresholds facilitates the introduction and application of new domain-specific concepts in teaching situations. As a general rule, learners must first develop an understanding of concepts along lower thresholds before being able to learn concepts along higher thresholds and expand their individual basis of knowledge and understanding. This study’s findings demonstrate that the participants had to first understand, for example, the difference between cash and profit before being able to learn concepts such as the break-even analysis. This hierarchical order should be further examined in future studies, for example, by analyzing so-called person-fit indices such as the hierarchy consistency index (e.g., Cui and Leighton 2009) in order to examine the extent to which learning along the thresholds is reflected in learners. Here, threshold characteristics could be further operationalized and analyzed using more detailed descriptions in order to determine whether learning follows a strict hierarchical process, or if and how previously “crossed” thresholds are reviewed by learners at a later stage (see also Davies 2012; Davies and Mangan 2007; Lucas and Mladenovic 2009).

When it comes to phrasing instructions, didactically appropriately selected scenarios should be used to devise different practice and application situations that allow for a better understanding of the concepts and that can be linked to subject-specific didactical principles as established in economics (e.g., problem orientation, case orientation, action orientation) (Böhner 2010). The application and practice of such concepts particularly requires an appropriate approach to dealing with learners’ understanding difficulties or errors (e.g., through scaffolding, see Oser and Spsychiger 2005). For learners, it is a great challenge to cross thresholds by developing a domain-specific understanding of the concepts rooted in these thresholds. This study’s findings directly indicate this phenomenon, as students have lower levels of familiarity with the concepts at higher thresholds and tend to apply guessing strategies more frequently. At the same time, guessing strategies are less effective the more challenging the concepts underlying the tasks are. In this context, it is crucial to create motivating learning environments in order to trigger and maintain learners’ motivation to learn these challenging concepts. Performance-oriented (video- or computer-based) teaching-and-learning tasks seem to have particularly

great potential in this regard (Butters and Walstad 2011; Oser et al., Chap. 7 in this volume) and should therefore be developed and used in the higher education sector. As recent studies show (e.g., Kuhn et al. 2016), such teaching-and-learning tools also offer promising development possibilities for formative and summative assessments in higher education.

In order to design the assessments, the threshold concepts' modeling approach needs to be considered when preparing the tests and tasks. The variety of task parameters – here, in the form of answer frequency – provides psychometric potential, as tests and tasks are specified through a combination of taxonomies, content areas, and threshold concepts, making it easier to predict test results. As mentioned earlier, the commonly used, traditional taxonomies alone are not sufficient as a basis for modeling and operationalization. As illustrated in a study by Kricks et al. (2013), the combination and variation of threshold, key and central concepts used in individual tasks provides promising potential for determining what learners know and understand. This is achieved by preparing tasks in exact accordance with learners' individual preconditions. The findings also emphasize that learning and test achievements cannot simply be predicted using threshold concept modeling. As made evident in the fluctuations of differences between how familiar a person is with a concept and how confident they are in completing a task, there also seem to be other (construct-irrelevant) characteristics that influence solution behavior and cannot be completely modeled and determined using threshold concepts alone (Kricks et al. 2013; Lucas and Mladenovic 2009). The increase in difference between confidence and familiarity with higher thresholds proves that other mental operations (e.g., test wiseness, Rogers and Yang 1996) are also significant.

The findings from this study show that a threshold concept approach can enhance the modeling of knowledge and understanding. Threshold concept modeling, however, should also encompass taxonomy modeling and should be examined in more detail. To achieve a solid assessment of the scope of these modeling combinations, more extensive studies with larger numbers of participants and tasks are required (e.g., Shanahan et al. 2006).

References

- Allgood, S., & Bayer, A. (2016). Measuring college learning in economics. In R. Arum, J. Roksa, & A. Cook (Eds.), *Improving quality in American higher education: Learning outcomes and assessments for the 21st century* (pp. 86–134). San Francisco: Jossey-Bass.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Asano, T., & Yamaoka, M. (2015). How to reason with economic concepts: Cognitive process of Japanese undergraduate students solving test items. *Studies in Higher Education*, 40(3), 412–436.
- Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning*. New York: Academic.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Handbook I: Cognitive domain. In *Taxonomy of educational objectives* (Vol. 19). New York: Longman.

- Böhner, M. M. (2010). Unterrichtsrelevante Prinzipien im Bereich Wirtschaft. *Erziehungswissenschaft und Beruf*, 58(3), 315–332.
- Brückner, S. (2017). *Prozessbezogene Validierung anhand von mentalen Operationen bei der Bearbeitung wirtschaftswissenschaftlicher Testaufgaben*. [Process-related validation using mental operations of responding to business and economics test items] Landau: Empirische Pädagogik.
- Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multi-level models to validate an assessment of higher education students' competency in business and economics. *Journal of Educational Measurement*, 53(3), 293–312.
- Butters, R. B., & Walstad, W. B. (2011). Computer versus paper testing in precollege economics. *The Journal of Economic Education*, 42(4), 366–374.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: L. Erlbaum Associates.
- Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, 46(4), 429–449.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *The American Psychologist*, 60(2), 170–180.
- Davies, P. (2012). Threshold concepts in economic education. In G. M. Hoyt & K. McGoldrick (Eds.), *International handbook on teaching and learning economics* (pp. 250–258). Cheltenham/Northampton: Edward Elgar.
- Davies, P., & Mangan, J. (2007). Threshold concepts and the integration of understanding in economics. *Studies in Higher Education*, 32(6), 711–726.
- Davies, P., & Mangan, J. (2009). *Understanding graphs in economics: An interpretation through threshold concepts*. Paper presented at the biennial conference of the European Association for Learning and Instruction (EARLI) in Amsterdam.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). Cambridge, MA: MIT Press.
- Gagné, R. M. (1985). *The conditions of learning and theory of instruction*. New York: CBS College Publishing.
- Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy. *The Journal of Educational Research*, 91(1), 26–32.
- Jang, K., Hahn, K., & Kim, K. (2010). Comparative Korean results of TUCE with U.S. and Japan. In M. Yamaoka, W. B. Walstad, M. Watts, T. Asano, & S. Abe (Eds.), *Comparative studies on economic education in Asia-Pacific region* (pp. 53–78). Tokyo: Shumpusha.
- Kricks, K., Mittelstädt, E., & Liening, A. (2013). Schwellenkonzepte und Phänomenografie: Explorative Studie zur Messung von Unterschieden im ökonomischen Verstehen [Threshold concepts and phenomenography: An explorative study on the measurement of differences in economic understanding]. *Zeitschrift für ökonomische Bildung*, 2, 17–41.
- Kuhn, C., Brückner, S., & Zlatkin-Troitschanskaia, O. (2016). *A new video-based assessment tool to enhance instructional practices of prospective teachers of economics*. Paper presented at the national conference on Teaching and Research on Economic Education (CTREE) in Atlanta.
- Leiser, D., & Aroch, R. (2009). Lay understanding of macroeconomic causation: The good begets-good heuristic. *Applied Psychology*, 58(3), 370–384.
- Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, 41(9), 352–362.
- Lucas, U., & Mladenovic, R. (2006). Developing new world views: Threshold concepts in introductory accounting. In J. Meyer & R. Land (Eds.), *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge* (pp. 148–160). London: Routledge.
- Lucas, U., & Mladenovic, R. (2009). The identification of variation in students' understandings of disciplinary concepts: The application of the SOLO taxonomy within introductory accounting. *Higher Education*, 58(2), 257–283.
- Lydersen, S., Pradhan, V., Senchaudhuri, P., & Laake, P. (2007). Choice of tests for association in small sample unordered r*c tables. *Statistics in Medicine*, 26, 4328–4343.
- Macha, K., & Schuhen, M. (2011). Framework of measuring economic competencies. *Journal of Social Science Education*, 10(3), 36–45.

- Meyer, J. H. F., & Land, R. (2003). *Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practising within the disciplines* (Occasional Report 4). Retrieved from ETL Project, Universities of Edinburgh, Coventry and Durham website: <https://kennslumid-stod.hi.is/wp-content/uploads/2016/04/meyerandland.pdf>
- Meyer, J. H. F., & Land, R. (2005). Threshold concepts and troublesome knowledge (2): Epistemological considerations and a conceptual framework for teaching and learning. *Higher Education*, 49(3), 373–388.
- Meyer, J. H. F., & Land, R. (2006). Threshold concepts and troublesome knowledge: An introduction. In J. Meyer & R. Land (Eds.), *Overcoming barriers to student understanding. Threshold concepts and troublesome knowledge* (pp. 3–18). London/New York: Routledge.
- Minnameier, G. (2013). The inferential construction of knowledge in the domain of business and economics. In K. Beck & O. Zlatkin-Troitschanskaia (Eds.), *Professional and VET learning: Vol. 2. From diagnostics to learning success. Proceedings in vocational education and training* (pp. 141–156). Rotterdam: Sense Publishers.
- Musekamp, F., & Pearce, J. (2016). Student motivation in low-stakes assessment contexts: An exploratory analysis in engineering mechanics. *Assessment & Evaluation in Higher Education*, 41(5), 750–769.
- O'Donnell, R. M. (2009). *Threshold concepts and their relevance to economics*. Paper presented at the 14th annual Australasian Teaching Economics Conference (ATEC 2009) (pp. 190–200). Brisbane, Queensland: School of Economics and Finance, Queensland University of Technology.
- Oser, F., & Spychiger, M. (2005). *Lernen ist schmerzhaft: Zur Theorie des negativen Wissens und zur Praxis der Fehlerkultur* [Learning is painful – On the theory of negative knowledge and a practice and error culture]. Weinheim: Beltz.
- Reimann, N., & Jackson, I. (2006). Threshold concepts in economics: A case study. In J. Meyer & R. Land (Eds.), *Overcoming barriers to student understanding. Threshold concepts and troublesome knowledge* (pp. 115–133). London/New York: Routledge.
- Riegler, P. (2014). Schwellenkonzepte, Konzeptwandel und die Krise der Mathematikausbildung [Threshold concepts, concept transition and the crisis of mathematical education]. *Zeitschrift für Hochschulentwicklung*, 9(4), 241–257.
- Rogers, W. T., & Yang, P. (1996). Test-Wiseness: Its nature and application. *European Journal of Psychological Assessment*, 12(3), 247–259.
- Schumann, S., & Eberle, F. (2011). Bedeutung und Verwendung schwierigkeitsbestimmender Aufgabenmerkmale für die Erfassung ökonomischer und beruflicher Kompetenzen. In U. Faßhauer, B. Fürstenau, & E. Wuttke (Eds.), *Schriftenreihe der Sektion Berufs- und Wirtschaftspädagogik der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE). Grundlagenforschung zum Dualen System und Kompetenzentwicklung in der Lehrerbildung [Übersetzung]* (pp. 77–90). Opladen: Budrich.
- Shanahan, M. P., Foster, G., & Meyer, J. H. (2006). Operationalising a threshold concept in economics: A pilot study using multiple choice questions on opportunity cost. *International Review of Economics Education*, 5(2), 29–57.
- Uribe, R. V. (2013). Measurement of learning outcomes in higher education: The case of Ceneval in Mexico. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education* (pp. 137–146). Rotterdam: Sense Publishers.
- Walstad, W. B. (2001). Improving assessment in university economics. *The Journal of Economic Education*, 32(3), 281–294.
- Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of understanding in college economics: Examiner's manual*. New York: National Council on Economic Education.
- Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from the German assessment of business and economics competence. In H. Coates (Ed.), *Assessing learning outcomes: Perspectives for quality improvement* (pp. 175–197). Frankfurt am Main: Lang.
- Zumbo, B. D. (2016). Standard-setting methodology: Establishing performance standards and setting cut-scores to assist score interpretation. *Applied Physiology, Nutrition, and Metabolism*, 41(6), 74–82.

Chapter 7

Rescue an Enterprise from Failure: An Innovative Assessment Tool for Simulated Performance



Fritz Oser, Susan Mueller, Tanja Obex, Thierry Volery,
and Richard J. Shavelson

Abstract Despite the fact that half of all start-ups fail during the first 5 years of their existence, failure is currently neglected in entrepreneurship education. We posit that what is needed is a competence that allows entrepreneurs to become aware of dangers and weaknesses in the firm – a kind of *Sense of Failure (SoF)* – and a competence that allows them to use heuristics to react to these dangers, i.e., the competence profile *Rescue an Enterprise from Failure (REF)*. In this paper, we discuss measures for capturing these two constructs and initial validation results. It is the first time that this kind of performance-oriented test instrument has been applied to measure an entrepreneurial competence supposed to prevent entrepreneurial failure. However, we want to point out that our test captures a “simulated” performance that requires participants to describe how they *would act* based on a written case.¹

¹Parts of this article are taken from our research report to the Swiss State Secretariat for Education, Research and Innovation (SERI): Volery, T. and Oser, F. (2016). Teilprojekt 2: Entwicklung von validen und reliablen Messinstrumenten für den “Sense of Failure” und “Sense of Success” im Gründungsprozess. Schlussbericht. St. Gallen/Fribourg.

F. Oser (✉) · T. Obex
University of Fribourg, Fribourg, Switzerland
e-mail: Fritz.Oser@unifr.ch; Tanja.Obex@unifr.ch

S. Mueller
University of St. Gallen, St. Gallen, Switzerland
e-mail: susan.mueller@unisg.ch

T. Volery
University of Western Australia, Perth, WA, Australia
e-mail: thierry.volery@uwa.edu.au

R. J. Shavelson
Stanford University, Stanford, CA, USA
e-mail: richs@stanford.edu

7.1 Entrepreneurial Dreams and a Blind Spot: Entrepreneurs' Failure

Most young entrepreneurs have a dream: they want to create something that has not been done before; they want to be their own boss, make their own decisions, and create wealth for themselves. Entrepreneurship education can be one way to instill this entrepreneurial spirit and increase entrepreneurial intention and capacity. However, entrepreneurship education can also have the opposite effect. Some students will – after participating in entrepreneurship training that allows them to experience what entrepreneurship entails – decide *not* to become entrepreneurs. This “sorting effect” can already be seen as an important result of an entrepreneurship education, since participants are able to make a more informed decision about their vocational choices.

Two other important functions of an entrepreneurship training program are that participants develop a positive knowledge base (how to do things) and a sense for potential failure. A great deal of research and development has focused on the former. However, participants must learn that failure is a possible and likely outcome of an entrepreneurial endeavor. Research shows that only one-third of all newly created firms worldwide will eventually develop into a profitable company (Reynolds 2016). This fact highlights the necessity to include entrepreneurial failure as an important topic in entrepreneurship education and to reflect on reasons, context conditions, and consequences of it. Course participants should develop a competence that increases the likelihood that the businesses they start in the future will survive.

Whether or not course participants develop such a competence depends on the content taught. Traditional entrepreneurship education programs include, for example, topics like the development of business ideas, the creation of a business plan, the conduct of market analyses, the creation of a marketing strategy, staff planning, and deciding on the legal structure of a start-up (e.g., see the textbook by Fueglistaller et al. 2016). What is missing in the majority of entrepreneurship textbooks and courses are stories, warnings, and knowledge pieces about failure.

As entrepreneurship educators, we therefore need to answer the question, “What do we need to include in our entrepreneurship training programs to avoid business failure or make business failure ‘less painful’?” How can we let students experience the consequences of specific mistakes? How can we allow them to develop a competence to detect potential dangers early on? Answering such questions is highly important when we want to equip program participants who opt for an entrepreneurial career with the necessary skill set to become successful entrepreneurs and – in case failure cannot be prevented – to avoid a potentially disastrous psychological and financial situation. In other words, in addition to other contents on entrepreneurial competencies, we need to develop something like a *Sense of Failure (SoF)*, a competence profile that allows students to become aware of potential dangers early enough, as well as a competence to *Rescue an Enterprise from Failure (REF)*. And

this must not only be reflected in the curriculum content but also in the instruments used to test the effectiveness of entrepreneurship education training programs.

The central claim of this paper is thus that we also need to include this possible negative side of the coin and acknowledge that failure is part of the entrepreneurship game.

7.2 Sense of Success and Sense of Failure: Goal Twins and Measurement Twins

Launching a new business venture necessitates a “can-do” attitude or entrepreneurial optimism; entrepreneurs must be convinced that they are able to develop an opportunity and that their offer will result in market acceptance (Rotefoss and Kolvereid 2005; Souitaris et al. 2007). In other words, entrepreneurs must have a *Sense of Success*. This has been acknowledged in both academic and practical terms. However, there are not many hints pointing to potential dangers in entrepreneurship (Bryant and Dunford 2008; Oser and Volery 2012).

We argue that this is an omission and that the competence *SoS* needs a “twin competence,” i.e., a *Sense of Failure (SoF)*, a prerequisite that allows individuals to minimize unnecessary breakdowns of young firms (Oser and Volery 2012). *SoF* is a prevention mechanism concerned with security, safety, and stability characteristics. It is a kind of spontaneous “seventh sense” which is responsible for applying the emergency brake at the right moment. *SoF* includes knowledge about potential pitfalls and is thus a precondition for the competence to deal with difficulties in all areas and at various stages of the start-up process.

This sense therefore captures the capacity to be sensitive to potential dangers, to be aware, and to remain alert to emerging problems, impasses, and micro-failures. Alertness – even if it has traditionally been used in economics as a capacity to be aware of new entrepreneurial opportunities – is a very appropriate dimension in this context. It draws on “the ability to notice something... without purposefully searching for it” (Kirzner 1979 in Frese and Gielnik 2014). In this respect, *SoF* is not about acting; it entails a highly developed capacity to find out something and display alertness in order to play a preventive role – in this case regarding entrepreneurial failure. *SoF* can be regarded as an instrument that influences the ability to recognize a problem and the ability to react.

The competence *SoF* includes items capturing *Negative Knowledge*, *Fear of Failure*, and responsibility attitudes and has been developed and validated throughout the course of a research project supported by the State Secretariat for Education, Research and Innovation (SERI) in Switzerland. Details of the construct and its validation process are provided in Oser and Obex (2015) and Oser and Volery (2012). Since *Negative Knowledge* plays a major role in the *SoF* competence, we will provide more insights into its theoretical basis in the following section.

7.3 Psychological Rootedness of Sense of Failure

The concept of failure is rooted in the psychological concept of *Negative Knowledge* (Hascher and Kaiser 2015; Oser and Spychiger 2005; Gartmeier and Schüttelkopf 2012; Harteis et al. 2012). *Negative Knowledge* refers to remembering events, things, procedures, or strategies that are *not* adequate, *not* effective, or even *false*. Remembering these issues is of high necessity for epistemic understanding. To know that a simple mathematical operation is correct means to know all the possibilities of its falseness. If someone knows what a money exchange rate is, *Negative Knowledge* is a kind of opposite index, hinting at what cannot be used as an exchange rate. If someone knows how to use the gear shift when accelerating a car, it is helpful to know how dangerous it is to shift into reverse when driving forward (for review, see Bauer and Harteis 2012; Gartmeier et al. 2015; Wuttke and Seifried 2012). If we apply this to a young firm, in order to protect a start-up from failure, it is important to know about other firms' mistakes (negative or failure knowledge), about how they recovered, and also about *almost-mistakes* (near misses). The concept of *almost-mistakes* captures the awareness that the entrepreneur – with his or her actions and decisions – just managed to avoid company failure (on *almost-mistakes*, see Oser et al. in “Human fallibility” 2012b).

Thus, avoiding serious errors through *Negative Knowledge* is an important quality of professional expertise. One explanation for expert performance is associated with the ability to avoid severe errors. A plausible but not yet widely considered explanation for such a capacity is the availability of explicit knowledge and skills about what not to do in certain situations. In this respect, avoiding or at least reducing individual mistakes is a necessary precondition of every outstanding success.

There are many reasons for entrepreneurial failure in the start-up phase of a company, including a lack of market interest for the products or services, personal or team conflicts, unnecessarily high fixed costs, liquidity issues, bad company strategies, and failed marketing measures. We do not think that all of these mistakes can be avoided, nor do we think that failure only has negative features. Nonetheless, there is also unnecessary failure, which can be prevented. And one should not forget that for young people failure can be a personal catastrophe. All of the few research projects that exist in this field (e.g., Cope 2011) show that failure goes hand in hand with loss of belief in oneself, the breakup of relationships, long periods of paying off debt, the stigma of being a loser, and other serious deficits.

7.4 Measuring Entrepreneurial “Sense of Failure”²

7.4.1 Preparing the New Instrument

Based partly on interviews with entrepreneurs whose enterprises failed and based on theoretical reflections and empirical findings, also relying on an exploratory study, we identified *Sense of Failure (SoF)* as a threefold construct with three dimensions: (1) *Negative Knowledge*, (2) *Fear of Failure*, and (3) *responsibility*. We were able to generate and validate items for each subdimension of *SoF*: 30 items to measure entrepreneurial *Negative Knowledge*, 29 items which represent different aspects of *Fear of Failure* (motivational vs. inhibitory function of fear, loss of control, elicitor for fear, sensitivity for fear), and 30 items to capture different aspects of the construct *responsibility* (dependency effects, authority reference, obligation and causality, hierarchy of responsibility).

The following steps have been undertaken to develop and validate the *Sense of Failure* construct.

Step 1: Generating Items of *Sense of Failure* – *Negative Knowledge*

To measure entrepreneurial *Negative Knowledge*, we formulated 30 items. Referring to the theory of *learning from mistakes* and the function of *Negative Knowledge* (Oser and Spychiger 2005), there are three sources of *Negative Knowledge*: (1) learning from one’s own experiences/mistakes, (2) learning from others’ experiences/mistakes, and (3) theoretical knowledge acquired in school, training, etc. For each of these sources, we formulated ten items.

Step 2: Generating Items of *Sense of Failure* – *Fear of Failure*

For measuring *Fear of Failure* in an entrepreneurial context, we generated 29 items which represent different aspects of fear (motivational vs. inhibitory function of fear, loss of control, elicitor for fear, goal orientation, performance orientation, and sensitivity for fear). Additionally, we used three factors (fear of having an uncertain future, fear of devaluing one’s self-estimate, fear of upsetting important others) with nine items on a scale for measuring entrepreneurial fear (Conroy 2001). This scale focuses on the consequences of failing; each factor is represented by three items.

²We are fully aware that if we measure entrepreneurial competencies in general, we must include knowledge and skills that are directly needed for a start-up. These are questionnaires about necessary tools, like being able to develop a business idea, make a business plan, conduct a market analysis, know about financing, have knowledge of legal matters, implement an advertising campaign, etc. We also need to measure the *Sense of Success* that contains entrepreneurial feasibility beliefs, entrepreneurial motivation, entrepreneurial risk-taking, entrepreneurial self-efficacy beliefs, striving toward professional autonomy, innovation affinity, entrepreneurial desirability, entrepreneurial stress resistance, entrepreneurial reliance, and entrepreneurial expectations (see Oser et al. 2012a with 22 such scales). However, in this chapter we want to stress the negative aspect of entrepreneurial failure.

Step 3: Generating Items of *Sense of Failure – Responsibility*

To operationalize *responsibility*, we formulated 30 items against a theoretical background (Noddings 2002; Sennett 2002; Jonas 1986) with respect to different aspects of the construct *responsibility* (dependency effects, authority reference, obligation and causality, hierarchy of responsibility, heuristics, ethics, and delegation of responsibility). In addition, we used a scale with 12 items to measure effectiveness and ethics in the context of enterprises (Tokarski 2008).

The instrument contains tasks where entrepreneurs value the extent of their agreement/disagreement. To examine the content validity of the instrument, the items were administered to six experts: two academics with expertise in entrepreneurship, one entrepreneur, and three academics with expertise in pedagogical psychology. We asked the experts to what extent does each item represent the intended construct. Based on their suggestions, we improved the wording of the items.

7.4.2 Quantitative Pre-study

Having developed the first version of the measurement instrument, we conducted a pre-study (paper-pencil) with economics students ($N = 109$) from Switzerland and Austria to reduce items and to elicit the structure of *SoF*. Even though we had an idea of the different sub-factors of the three constructs, exploratory factor analyses were conducted in this pre-study; the goal was to get a picture of the correspondence of the items and to discard items.

By doing exploratory factor analyses and item analyses, we reduced the items for *Fear of Failure* (from 21 to 15 items) and identified five dimensions of *entrepreneurial Fear of Failure: elicitor, consequences, inhibitory function, stimulating function, and manifest fear*. A further exploratory factor analysis of *responsibility* did not lead to a solution that is interpretable in relation to its content. After reducing the items due to verbal content, the *responsibility* scale (32 items) was modified and now consists of eight items. Furthermore, we eliminated Tokarski's (2008) scale for measuring effectiveness and ethics because of the low alpha coefficients of the proposed factors (.464–.630) and the proposed factor structure's insufficient model fit. Due to the amount of missing data, we could not analyze *Negative Knowledge*.

7.4.3 Structural Validity Study

Building on the pre-study, we conducted an online survey to establish the structural validity of the proposed instrument. The first sampling frame consisted of 2048 entrepreneurs and managers of small- and medium-sized enterprises (SMEs),

mainly from Switzerland, who have taken part in some SME training programs over the past 10 years. After sending an email invitation to all listed individuals, we discovered that 155 had invalid email addresses, leaving us with a potential sampling frame of 1893 potential respondents. Following the first email, we sent two reminder invitations at weekly intervals. After 3 weeks, we closed the survey and examined the responses to sift out incomplete surveys and the few that exhibited suspicious response patterns (like providing the same answer to a series of unrelated items). This left us with 232 valid surveys (giving an effective response rate of about 12.2%).

7.4.4 *Fear of Failure*

To investigate the structure of the improved scale capturing *Fear of Failure*, we conducted an exploratory factor analysis with principal component analyses and promax rotation; 15 items were included. Due to lack of interpretability, the results of a first item analysis prompted us to discard five items. This decision was related to the content and to low reliability and low item-to-total correlations. An additional exploratory factor analysis with nine items revealed a three-factor solution (see Table 7.1).

The Kaiser-Meyer-Olkin (KMO) was calculated to be 0.794, a mediocre fit. The factor loadings of the items range from .648 to .877, which suggests a reasonable three-factor solution accounting for 65.83% of the total variance. To determine the internal consistency reliability, Cronbach's alpha values were calculated: factor 1 (*manifest fear*) contains five items, and its reliability coefficient is .797; factor 2 (*stimulating function of fear*) contains three items which has a reliability coefficient of .661; and factor 3 (*inhibitory function of fear*) contains two items producing an alpha value of .695. The reliability coefficient of the complete scale for *Fear of Failure* – assuming a single factor – is .757. We obtain item-to-total correlations between .215 and .557 (average = .433).

To check the results of the exploratory factor analyses, the structure of *Fear of Failure* was verified in a second step applying a confirmatory factor analysis (CFA). Using a data set without missing values ($N = 231$), a model (1) including three latent factors (*manifest Fear of Failure*, *stimulating function*, *inhibitory function*) was tested. This three-factor model was compared to a single construct (2) *Fear of Failure*. We estimated several fit indices (CFI, TLI, and RMSEA) to evaluate the steadiness of our model. The results indicated that the fit of model 1 ($\chi^2_{(24)} = 27.309$, $p = .290$; TLI = .990; RMSEA = .024) was better than the single-construct model 2 ($\chi^2_{(27)} = 137.891$, $p = .000$; TLI = .712; RMSEA = .134). Table 7.2 illustrates the multiple squared correlations of each item and indicates sufficient indicator reliability.

Table 7.1 Results of EFA and item analyses for *Fear of Failure* (NEW)

	Factor 1	Factor 2	Factor 3
	Manifest Fear of Failure	Stimulating function of fear	Inhibitory function of fear
When I think of failure, my heartbeat quickens	.792	.244	.236
When something is not going well in the company, I immediately get frightened	.780	.313	.210
I am sometimes so worried about my company that my hands tremble	.775	.231	.106
Just thinking about failure makes me often miserable in the morning	.750	.355	.123
Even if I work a lot and do many overtime hours, I think about what could go wrong	.648	.497	.140
Fear of Failure helps me to do a lot of things in a better way	.269	.877	-.051
Fear of Failure drives me to work even more diligently	.399	.828	.159
If I am afraid of my business, I cannot make any decisions	.193	.061	.874
If I am afraid of my business, I cannot work well	.177	.032	.872
Variance explained	37.12%	16.59%	12.12%
Cronbach's α (per factor)	.797	.661	.695
KMO	.794		
Variance explained (total)	65.83%		
Cronbach's α (total)	.757		
Composite reliability	.775		

Table 7.2 Squared multiple correlations of the items of *Fear of Failure*

	Squared multiple correlations
When I think of failure, my heartbeat quickens	.513
When something is not going well in the company, I immediately get frightened	.511
I am sometimes so worried about my company that my hands tremble	.453
Just thinking about failure makes me often miserable in the morning	.459
Even if I work a lot and do many overtime hours, I think about what could go wrong	.359
Fear of Failure helps me to do a lot of things in a better way	.363
Fear of Failure drives me to work even more diligently	.675
If I am afraid of my business, I cannot make any decisions	.576
If I am afraid of my business, I cannot work well	.493

7.4.5 Responsibility

The reliability coefficient of the *responsibility* scale is .592, with a respective factor reliability of .605 (see Table 7.3). We obtain item-to-total correlations between .309 and .418 (average = .333). The exploratory factor analyses conducted detected factor loadings between .534 and .668; those six items explain 33.64% of the variance.

Table 7.3 Results of the EFA, CFA, and item analyses of *responsibility*

	EFA	CFA
	Factor loading	Squared multiple correlations
If an entrepreneur fails, everything is lost; minimal damage is rather not possible	.668	.323
If a start-up goes bankrupt, there are external reasons for that (e.g., an economic crisis). The founders can do nothing about this event. They cannot predict it	.603	.223
This mostly happens by chance	.568	.181
If my firm crashes, I have to think about me first. I have to think about <i>my</i> future. Ultimately, it is me who has invested	.551	.183
In case of failure, the responsibility of a start-up must be directed toward economic goods and only then toward the people concerned	.545	.170
A real founder has no responsibility to society	.534	.155
KMO	.714	
Variance explained	33.64%	
Composite reliability (CFA in Amos)	.605	
Cronbach's α	.592	

Even though the squared multiple correlations of a conducted confirmatory factor analysis are low, the suggested factor structure has a good model fit ($\chi^2(9) = 11.988$, $p = .214$; TLI = .950; RMSEA = .038).

7.4.6 Negative Knowledge

To investigate the structure of *Negative Knowledge* responses, an exploratory factor analysis was applied with principal component analyses and promax rotation; 30 items were included. The results recommended a nine-factor solution that is not interpretable in relation to its content. Consequently, we decided to exclude the sources of *Negative Knowledge* and to focus on the *functions of Negative Knowledge*.

After a series of exploratory factor analyses (EFA), we excluded 24 items and decided the one-factor solution with six items was most interpretable. The one-factor solution explained 42.36% of the variance (see Table 7.4). The factor loadings of the items range from .429 to .730, which is reasonable. This solution was examined in a confirmatory factor analysis, and a good model fit was achieved ($\chi^2(9) = 12.140$,

$p = .206$; $TLI = .975$; $RMSEA = .040$). The squared multiple correlations of the single items were, with the exception of one item, acceptable to good (from .237 to .444). An item analysis shows item-to-total correlations from .204 to .432 (average = .316), and the alpha reliability is .712 (composite reliability: .726).

Table 7.4 Results of the EFA, CFA, and item analyses of *Negative Knowledge*

	EFA	CFA
	Factor loading	Squared multiple correlation
Those who recover from failure know the sources of mistakes in a start-up better	.730	.444
Failure prevents certain start-up mistakes from reoccurring	.722	.436
Those who are able to overcome a crisis in the founding process are more aware of start-up mistakes	.684	.341
We learn from failure	.679	.336
You learn the most when the success of a company is standing on the brink	.612	.237
Anyone who fails with a company learns how painful failure can be	.429	.100
KMO	.789	
Variance explained	42.36%	
Composite reliability	.726	
Cronbach's α	.712	

7.4.7 Sense of Failure As a Construct

Several exploratory factor analyses, item analyses, and confirmatory factor analyses left us with 21 candidate items and three dimensions of *SoF*: six items for *responsibility*, six items for *Negative Knowledge*, and nine items split into three factors for *Fear of Failure*. Putting these factors together with the goal of achieving a model with a higher-order factor did not work. As a consequence, we discarded the factors stimulating *Fear of Failure* and inhibiting *Fear of Failure*.³ With these 3 remaining dimensions and 11 items, we ran a confirmatory factor analysis (CFA) to examine the structural validity of *SoF* as a higher-order factor (Fig. 7.1).

The results indicate a good model fit ($\chi^2(118) = 139.19$, $p = .089$; $TLI = .961$; $RMSEA = .029$) for *SoF* as one construct including three dimensions: (1) *Negative Knowledge*, (2) *responsibility*, and (3) *manifest fear*. The dimensions *manifest fear*, *Negative Knowledge*, and *responsibility* are uncorrelated (correlation coefficients range from $-.117$ to $.052$). These results indicate that *SoF* is a multidimensional construct that is constituted by the factors *Negative Knowledge*, *responsibility*, and *manifest Fear of Failure*.

³CFA for *manifest fear*: $\chi^2_{(5)} = 6.753$, $p = .240$; $TLI = .989$; $RMSEA = .040$.

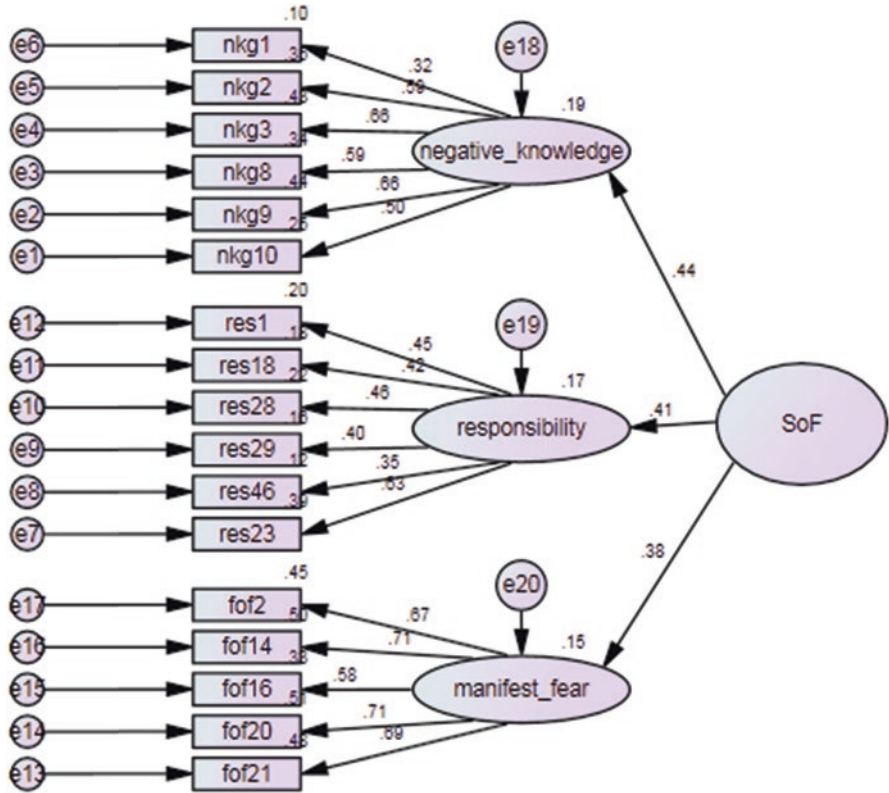


Fig. 7.1 Confirmatory factor analysis for *Sense of Failure*

7.5 Rescue an Enterprise from Failure (REF): An Emergency Capability

The *REF* competence profile is an emergency capability. It consists of adaptive professional acts, which are validated through performances and leads to goal-oriented change in the state of a person or the state of a system. The competence profile makes it possible to consider alternative courses of action instead of merely following routinized forms of actions that are often carried out in an uncontrolled manner and without reflection.

To measure the *REF* competence, we opted to use a *performance assessment* approach. This concept was explained by Shavelson (2013; see also Shavelson et al., Chap. 10 in this volume). Obviously, performance assessment is only one of various methods of gathering evidence for the competence of a person. We accept that even numerous performances do not fully capture a specific competence (internal validity gap). Nonetheless, the performance of an action represents a high-fidelity attempt to capture the underlying competence. In extremis, it is possible that

someone may successfully perform an action by chance, but most professional competencies are of such complexity that an adequate performance by chance alone is not very likely. In short, we understand that the performance of a competence is a necessary but not sufficient condition for modeling that competence. These basics are discussed elsewhere in more detail (Oser 2013).

The *REF* competence profile that is needed for entrepreneurs to address adverse start-up situations consists of four capacities (see sketch in Fig. 7.2):

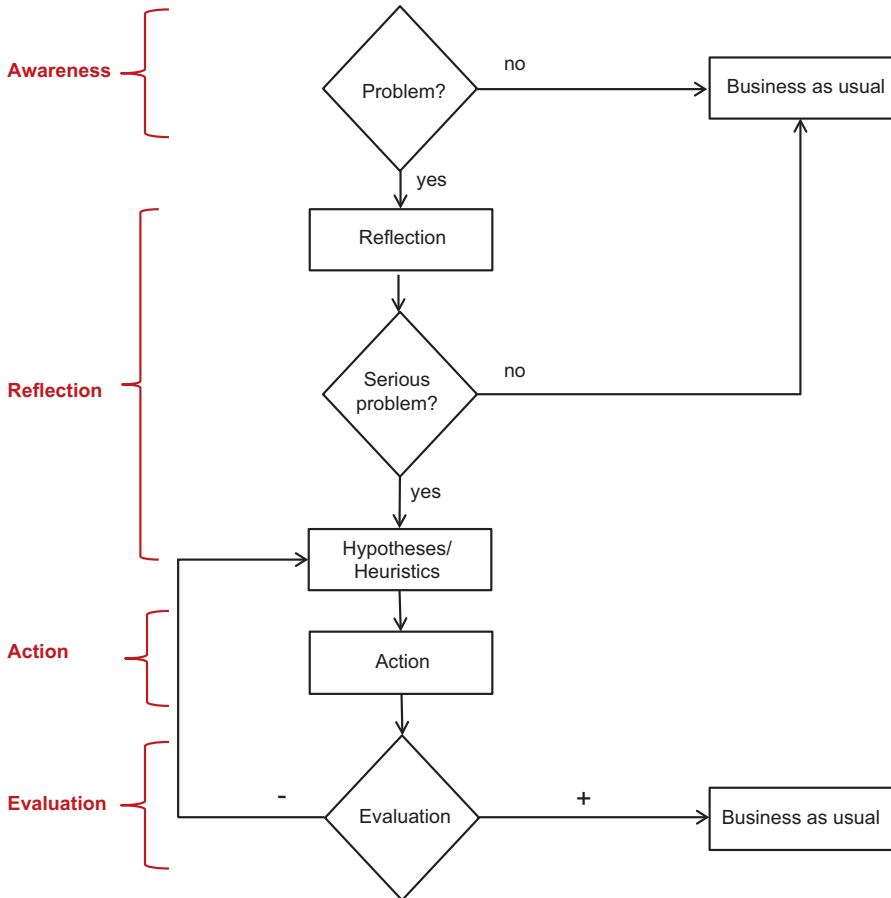


Fig. 7.2 The *Rescue an Enterprise from Failure* competence profile: a four-step model

- *Awareness*: In the first step, it is necessary to become aware that something could or is going wrong. *SoF* positively influences this capacity. Aside from becoming aware of an emerging problem, it is essential that work and processes are interrupted. This pause is a necessary condition for the following steps and implies an understanding of the situation as a necessity to do something.
- *Reflection*: This step is about reflecting on a situation and deciding whether a problem is serious or not. If the situation is considered to be serious enough, further steps have to be taken. Furthermore, the causes of the respective situation have to be explored, and potential actions to remedy/rescue the situation have to be reflected upon. The person needs to be able to develop hypotheses about possible causes for the situation. In addition, the capacity is about being able to give reasonable weight of importance to different information sources and being able to judge the reliability, relevance, or both of available information sources.
- *Action*: A particular action is chosen as a result of reflection. This partial competence refers to making a decision and doing something to change the situation at hand (the weaknesses, the errors, and the downfall). This step is therefore about making a global decision to substantially change the situation. (It should be noted that as part of our measurement instrument, the respondents do not implement the action but write about the potential actions they would take to improve the situation.)
- *Evaluation*: In this step, questions related to the success of the course of action are predominant. One has to find out if the action taken worked and why it worked. The result of such considerations either leads to continuation of the work process (“business as usual”) or initiates new reflection. The competence parts are (a) to evaluate one’s own decision by comparing and evaluating it with other peoples’ decision cluster and (b) to have the capability to restart or to adapt fully to the decision made. In Fig. 7.2, we tried to sketch the step-by-step procedure, here of course generalizing the respective lines to follow.

Awareness To capture this part competence, an entrepreneurial situation is presented to the respondents as a central stimulus, without *explicitly* mentioning the dangerous part of the situation. The situation describes a young firm that produces a lifestyle drink that is distributed by a major retail chain. In the story, the founders decided against self-distribution in bars, cafes, and so on and instead committed to exclusively supply a retail chain for 3 years with their lifestyle drink. Sales started well, and the founders bought a new bottling plant financed by a bank loan.

We then present seven potential dangers of the situation and ask respondents to evaluate them on a four-point scale (Fig. 7.3).

What are potential sources of danger? Please decide for each aspect whether it represents a danger (1= no danger, 4 = danger)				
	No danger			Danger
	1	2	3	4
Exclusive contract with a retail chain	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Purchase of the new bottling machine	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Three-year commitment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Bank credit	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Composition of the product	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The product itself –similarity to other energy drinks	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Problem with the two other retail chains	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Fig. 7.3 REF's operationalization of "awareness"

Calculating an Individual Score for Awareness To determine one's awareness, we calculated a mean from the points which the respondents received for their assessments. Respondents received 2 points if they correctly identified the first four items as a danger (and choose 3 or 4 on the scale) and 2 points if they recognized that the last three items did not represent a danger (and choose 1 or 2 on the scale).

Reflection By introducing new information into the situation, we intended to make the situation more complex. We therefore introduced the following problem: the retail chain wants to withdraw from the contract due to damaging circumstances. "Cola" threatened the retail chain that it would stop supplying its products if the retail chain did not stop selling the young company's lifestyle drink.

We ask the respondents (1) if they evaluated this problem as serious and if a reaction is necessary, and (2) we presented eight possible reactions and asked how useful each was on a four-point scale (Fig. 7.4).

Several possible actions are listed below. Please decide for each one how reasonable you think it is (1=not reasonable, 4 = very reasonable)				
<i>To rectify the situation, it would be possible to...</i>	<i>Not reasonable</i>		<i>Very reasonable</i>	
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
1. Engage a lawyer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Conclude a contract with another retail chain	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Seek a personal conversation with the people responsible at the retail chain	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Charge Coca Cola for reputational damage	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Sell everything, sell the recipe to pay back the debts and maybe still make a profit	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Make use of free legal advice	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Do nothing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. Sell the drink through cafes, bars, clubs etc.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Fig. 7.4 REF’s operationalization of “reflection”

Afterward, test-takers were provided with documents that offered different, potentially conflicting information (see Fig. 7.5: letter from a lawyer, letter to the editor published in a newspaper, document from the Internet including information about potential dangers of energy drinks, an email from a friend, a suggestion of a colleague in a team meeting, an advertisement of a new energy drink, a public invitation to participate in an award for young entrepreneurs).

Some documents are relevant to the situation and some not, and some are reliable and some are not (see Fig. 7.6). The goal was to constitute a real-life situation in which coping with different information is crucial. To ensure that people read this additional information, the items depicted in Fig. 7.6 have to be answered.

Calculating Individual Score for Reflection First, the research team evaluated each course of action according to its appropriateness in the given context. Two items, which were evaluated as inappropriate courses of action, were reversed in the data analysis.

To determine a person’s ability to find rescue heuristics, we calculated a mean score for the items being seen as useful. The theoretical presumption is that the more alternatives mentioned as useful, the higher the capability for further steps. To

evaluate the result of the task in which respondents had to assess additional documents, we gave one point for each correct assessment.

Action In this step, we ask respondents about their particular course of action in this situation. They must be able to consider all relevant and reliable information and the aforementioned rescue strategies, exclude irrelevant and unreliable information, and decide on their individual resolution strategy. Respondents were free to generate as many possible courses of action as they wished.

Calculating Individual Score for Action The alternatives given by the respondents were analyzed on the basis of three categories, and a total score was calculated:

1. Number of alternatives: for one alternative a person gets 1 point, and for two or more alternatives a score of 2 points is given.

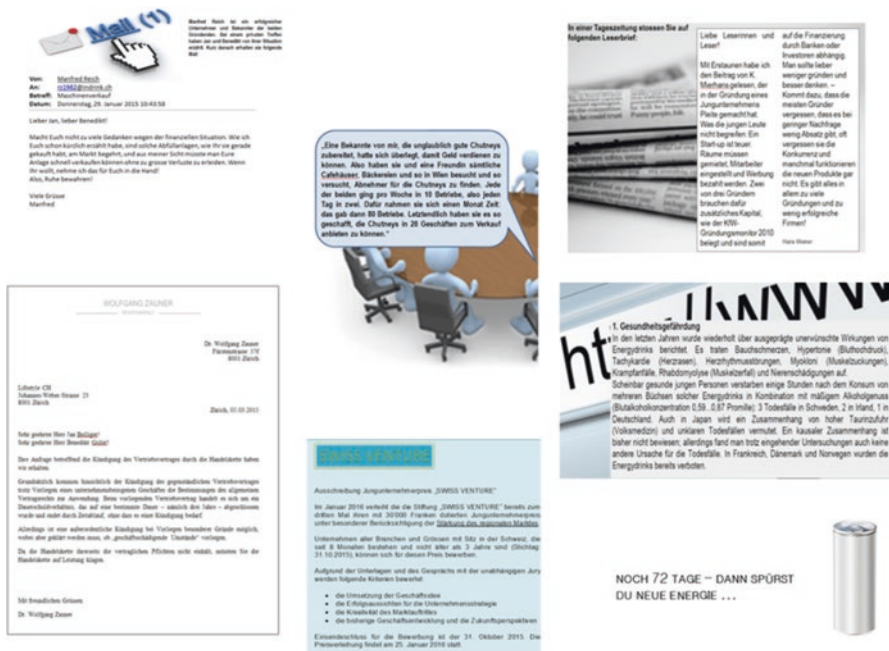


Fig. 7.5 REF's additional information

	<i>reliable</i>		<i>relevant</i>	
	<i>yes</i>	<i>no</i>	<i>yes</i>	<i>no</i>
Letter to the editor published in a newspaper		X		X
Advertisement of a new energy drink		X		X
Mail from a friend		X	X	
Discussion in a team-meeting		X	X	
www document / information about potential dangers of energy drinks	X			X
Public invitation / young entrepreneurs' prize	X		X	
Letter from a lawyer	X		X	

Fig. 7.6 Reliability and relevance of the additional information

2. Time or relevance order: we evaluated whether the given answers included a time order or a hierarchy; if this condition is fulfilled, the entrepreneur received 1 point, otherwise 0 point.
3. Quality: if the proposed alternative included a legitimate justification or if the alternative was judged a creative solution, the respondent received 2 points, otherwise 1 point.

To get a total score, we used the following formula: $(number + order) \times quality$. Scores ranged from 0 to 6.

Evaluation In order to explore respondents’ ability to evaluate their own decisions, we present them with four possible actions together with the consequences of each action and asked them how useful they deemed the particular alternatives to be on a four-point scale. One of the alternatives is shown in Fig. 7.7.

Alternative 1

The founders decided to distribute the drink through bars, cafes, and clubs and within a short amount of time they found 27 locations that will offer the drink. In total, they can now expect to sell 40,000 bottles per month for the next 18 months.

Please assess how reasonable this course of actions was.

Not reasonable Rather not reasonable Rather reasonable Very reasonable

Fig. 7.7 REF operationalization of “evaluation”

Calculating an Individual Score for Evaluation As described and applied in the reflection step, we also evaluated each possible action according to its appropriateness. Actions that were evaluated as inappropriate (alternative 2, alternative 3) were reversed in the data analysis. To determine a person's ability to evaluate specific courses of action, we calculated a mean score.

To determine *REF*'s feasibility, two experienced entrepreneurs first went through the whole testing process and subsequently gave us advice for clarification. Based on their advice, we modified the wording of the scenario and the additional information and improved the items. Then, four entrepreneurs used the whole *REF* for the first time in a pilot test. They all stated (a) that the test is too long. This prompted us to drop the idea of using a possible second scenario in the main study and to substantially reduce the qualitative parts of the questionnaire. They said that (b) the test was helpful and that many situations in their own enterprise reminded them of similar reactions and complex rescue trials.

7.6 The Connection Between Sense of Failure and the Competence to Rescue an Enterprise from Failure

After having developed the *SoF* and *REF* instruments, we conducted a final study to investigate the relationship of *SoF* on one hand and *REF* on the other hand. Through face-to-face interviews using a structured questionnaire and an additional online survey, we received 77 valid responses. To extend validity, we sought a mix of entrepreneurs with regard to (1) industry backgrounds, (2) early versus later stage, (3) with and without a failure experience, and (4) fast versus slow-paced industries.

According to the four-step model of *REF*, which represents four dimensions of the competence, Table 7.5 presents sample sizes, means, standard deviations, as well as minimum and maximum values for all part competencies, for each step in *REF*.

Entrepreneurs did not evaluate the proposed dangers as dramatic in the specific situation (mean = 2.50, SD = .496), while they interpreted the suggested options to react in this specific case as appropriate (mean = 3.48, SD = .375).

For competence "action," possible points ranged from 0 to 6, and we calculated an average score of 2.55 (SD = 1.77). The described action alternatives (including consequences and a new description of the situation) were evaluated on a four-point usefulness scale; a mean score of 2.93 (SD = .473) was calculated. Histograms of the part competencies showed that "awareness," "reflection," and "evaluation" are approximately symmetrically distributed. The part competence "action" deviates from a symmetrical distribution (kurtosis = -.264).

Table 7.6 provides the correlation coefficients between the *REF* total score and the part competencies of *REF*. The first column presents part-whole correlations. They show how much each of the four sub-competencies contributes to the total score. We note that the sub-competencies are uncorrelated.

Table 7.5 Statistical parameters of each part competence of REF

	<i>n</i>	Mean	SD	Variance	Min	Max
Awareness*	75	2.50	.496	.247	1.00	3.60
Reflection*	76	3.48	.375	.141	2.40	4.80
Action**	60	2.55	1.770	3.133	0.00	6.00
Evaluation*	67	2.93	.473	.224	1.50	4.00

*Four-point Likert scale; **score from 0 to 6

Table 7.6 Correlations between the REF total and the part competencies

	REF total	Awareness	Reflection	Action
Awareness	.698**			
Reflection	.519**	.080		
Action	.565**	.088	.122	
Evaluation	.486**	.219†	.020	-.012

Note: Significance levels: ** $p < .01$; † $p < .10$

Table 7.7 Regression analyses for SoF and responsibility predicting REF and part competencies of REF

Independent variable	Dependent variable				
	REF	Awareness	Reflection	Action	Evaluation
(Constant)	-6.882	1.272	2.514	4.535	2.142
SoF	2.495	.445	.347	-	-
Responsibility	-	-	-	1.146	.237
R^2	.089	.055	.057	.102	.054
F	4.702*	4.152*	4.445*	6.492*	3.650†

Note: $N = 77$

Significance levels: * $p < .05$; † $p < .10$

Correlation analyses exhibited low to moderate connections between SoF and REF and the dimensions of SoF and the part competencies of REF. We also wanted to get an idea of whether SoF and one or more of the elements of SoF might be the focus of training in schools to enhance one’s REF. The results of linear regression of SoF on total REF and each part competence of REF (awareness, reflection, action, evaluation) are shown below (Table 7.7). The construct SoF predicts almost 9% of one’s competence to *Rescue an Enterprise from Failure*. A person high on SoF – constituted by *Negative Knowledge*, *responsibility*, and *Fear of Failure* – tends to have higher competence in tackling an adverse situation in an entrepreneurial context (but the relationship is not strong). We note that *responsibility* did not predict REF.

SoF also has a statistically significant effect ($F = 4.152$; $p < .05$) on the part competence awareness and explains approximately 6% of the variance. The higher the SoF, the higher a person’s *Negative Knowledge*, *responsibility*, and *manifest fear* (physical symptoms of fear like sleep disorders, excitement, etc.), i.e., the higher one’s awareness of potential dangers, the more an entrepreneur is sensitive to criti-

cal situations in an entrepreneurial context. We note that *responsibility* did not predict awareness.

The regression of *SoF* on reflection shows that *SoF* influences entrepreneurs' reflection scores, explaining almost 6% of the variance. Again, this is a weak relationship. In contrast to awareness and reflection, it is not *SoF* but responsibility that explains 10% of the variance in predicting a person's ability to act. The regression predicting the ability to evaluate a specific course of action *from* responsibility shows a very weak relationship (less than 6%). The positive relationship between *SoF* and a person's competence to tackle adverse situation in entrepreneurial context (*REF*) suggests that it might be worthwhile to study the relation of *SoF* in schools or as part of vocational training to see if *REF* can be enhanced. This result begs the question of how specifically *manifest Fear of Failure* can be worked through and if it is ethically justified to sensitize a person to *Fear of Failure*. Furthermore, our study revealed that entrepreneurs with a higher *responsibility* tended to achieve higher values in processing tasks of action and evaluation in *REF*.

Consequently, we will plan an intervention study to foster young people's *SoF* and especially *responsibility* to enhance their ability to detect and reflect on critical situations in an entrepreneurial context on the one hand, and the quality of their actions and ability to evaluate a chosen course of action on the other hand. It is also necessary to focus on *Fear of Failure* and to elicit what effects *Fear of Failure* demonstrates.

7.7 Conclusion

Entrepreneurs need to be optimistic and courageous. They need to believe in their ideas. However, it is not enough to have high confidence; it is also necessary to develop a relevant *SoF* and a competence to act fruitfully in adverse situations in order to avoid failure or at least to fail in a way that does not "hurt" as much and does not leave the entrepreneur financially ruined.

We developed the competence profile *Rescue an Enterprise from Failure* on the basis of the theory of performance testing (Shavelson 2012; see also Shavelson et al., Chap. 10 in this volume). We designed its precondition, *Sense of Failure*, on the basis of literature reviews and qualitative interviews with persons whose companies failed. Both concepts can potentially help to reduce entrepreneurial failure. Currently, we are using both instruments for an ongoing intervention study at vocational schools in Switzerland, which aims to increase participants' *Sense of Failure* and their *REF* competence profile.

We are aware that failure will always be part of entrepreneurial activity, specifically in the case of truly innovative start-ups in the Schumpeterian understanding. However, in many cases, entrepreneurial failures are preventable or could at least be recognized earlier, meaning the consequences could be less harmful for the people involved. We are also aware that emphasizing failure in entrepreneurship education is a dilemma. Often, the objective of educators and politicians is to encourage more young people to become entrepreneurs. Highlighting potential dangers and negative

consequences thus could be counterproductive. However, responsible entrepreneurship education entails equipping students with the necessary set of skills to identify and exploit an entrepreneurial opportunity while at the same time protecting them against the hardship of failure.

We note that these are preliminary results. It is highly probable that the final measurement instruments will look different, since we will collect much more data in the current intervention study. This will give us the opportunity to further refine the instrument. Nonetheless, what we want to present is the idea of how teachers can positively treat negative events (failure scenarios) and how it is possible to measure the effects.

Indeed, we suggest that fostering young people's *Sense of Failure* and *Rescue an Enterprise from Failure* competencies may on the one hand enhance their ability to detect potential dangers and reflect on critical situations in an entrepreneurial context, and enhance their ability to evaluate an adequate course of action to address the situation on the other hand. We believe that our approach has no inhibiting effect on entrepreneurial intention but rather equips course participants with a protective device that might reduce entrepreneurial failure.

References

- Bauer, J., & Harteis, C. (Hrsg.). (2012). *Human fallibility: The ambiguity of errors for work and learning*. Dordrecht: Springer.
- Bryant, P., & Dunford, R. (2008). The influence of regulatory focus on risky decision-making. *Applied Psychology, 57*(2), 335–359.
- Conroy, D. E. (2001). Progress in the development of a multidimensional measure of fear of failure: The Performance Failure Appraisal Inventory (PFAI). *Anxiety, Stress, & Coping, 14*, 431–452.
- Cope, J. (2011). Entrepreneurial learning from failure: An interpretative phenomenological analysis. *Journal of Business Venturing, 26*(6), 604–623.
- Frese, M., & Gielnik, M. M. (2014). The psychology of entrepreneurship. *Annual Review of Organizational Psychology and Organizational Behavior, 1*(1), 413–438.
- Fueglistaller, U., Müller, C., Müller, S., & Volery, T. (2016). *Entrepreneurship Modelle – Umsetzung – Perspektiven Mit Fallbeispielen aus Deutschland, Österreich und der Schweiz* (4th ed.). Wiesbaden: Springer Gabler.
- Gartmeier, M., & Schüttelkopf, E. (2012). Tracing outcomes of learning from errors on the level of knowledge. In J. Bauer & C. Harteis (Hrsg.), *Human fallibility: The ambiguity of errors for work and learning* (S. 33–51). Dordrecht: Springer. DOI: https://doi.org/10.1007/978-90-481-3941-5_3
- Gartmeier, M., Gruber, H., Hascher, T., & Heid, H. (Hrsg.). (2015). *Fehler – Ihre Funktionen im Kontext individueller und gesellschaftlicher Entwicklung / Errors – Their functions in context of individual and societal development*. Münster: Waxmann.
- Harteis, C., Bauer, J., & Heid, H. (2012). Research on human fallibility and learning from errors at work: Challenges for theory, research, and practice. In J. Bauer & C. Harteis (Eds.), *Human fallibility: The ambiguity of errors for work and learning* (pp. S.255–S.265). Dordrecht: Springer.
- Hascher, T., & Kaiser, C. (2015). The acquisition of negative knowledge during field experience in teacher education. In M. Gartmeier, H. Gruber, T. Hascher, & H. Heid (Hrsg.), *Fehler – Ihre Funktionen im Kontext individueller und gesellschaftlicher Entwicklung / Errors – Their functions in context of individual and societal development* (S. 227–244). Münster: Waxmann.

- Jonas, H. (1986). *Das Prinzip der Verantwortung. Versuch einer Ethik für die technische Zivilisation* (*The principle of responsibility. Towards an ethic for the technical civilisation*). Frankfurt: Insel.
- Kirzner, I. (1979). *Perception, opportunity and profit: Studies in the theory of entrepreneurship*. Chicago: University of Chicago Press.
- Noddings, N. (2002). *Starting at home. Caring and social policy*. Berkley: University of California Press.
- Oser, F. (2013). "I know how to do it, but I Can't do it": Modeling competence profiles for future teachers and trainers. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education: Tasks and challenges* (pp. 45–60). Rotterdam: Sense Publishers.
- Oser, F., & Obex, T. (2015). Gains and losses of control: The construct "sense of failure" and the competence to "rescue an enterprise from failure". *Empirical Research in Vocational Education and Training*, 7(3), 1–17.
- Oser, F., & Spychiger, M. (2005). *Lernen ist schmerzhaft. Zur Theorie des Negativen Wissens und zur Praxis der Fehlerkultur*. Weinheim: Beltz Verlag.
- Oser, F., & Volery, T. (2012). "Sense of failure" and "sense of success" among entrepreneurs: The identification and promotion of neglected twin entrepreneurial competencies. *Empirical Research in Vocational Education and Training*, 4(1), 27–44.
- Oser, F., del Rey, N., Näpflin, C., Mosimann, S., Volery, T., & Müller, S. (2012a). *Abschlussbericht Entrepreneurship Programm: "Initiative Zukunft". Eine Interventionsstudie zur Erhöhung des unternehmerischen Kompetenzprofils bei Lernenden der Sekundarstufe II*. St. Gallen/Freiburg: KMU-HSG, Universität St. Gallen/Departement Erziehungswissenschaften, Universität Freiburg.
- Oser, F., Näpflin, C., Hofer, C., & Aerni, P. (2012b). Towards a theory of negative knowledge (NK): Almost-mistakes as drivers of episodic memory amplification. In J. Bauer & C. Harteis (Hrsg.), *Human fallibility: The ambiguity of errors for work and learning* (S. 53–70). Dordrecht: Springer.
- Reynolds, P. D. (2016). Start-up actions and outcomes: What entrepreneurs do to reach profitability. *Foundations and Trends® in Entrepreneurship*, 12(6), 443–559.
- Rotefoss, B., & Kolvereid, L. (2005). Aspiring, nascent and fledgling entrepreneurs: An investigation of the business start-up process. *Entrepreneurship & Regional Development*, 17(2), 109–127.
- Sennett, R. (2002). *Respekt im Zeitalter der Ungleichheit*. Berlin: Berlin Verlag.
- Shavelson, R. J. (2012). Assessing business-planning competence using the Collegiate Learning Assessment as a prototype. *Empirical Research in Vocational Education and Training*, 4, 77–90.
- Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist*, 48(2), 73–86.
- Souitaris, V., Zerbini, S., & Al-Laham, A. (2007). Do entrepreneurship Programmes raise entrepreneurial intention of science and engineering students? The effect of learning, inspiration and resources. *Journal of Business Venturing*, 22(4), 566–591.
- Tokarski, K. O. (2008). *Ethik und Entrepreneurship. Eine theoretische sowie empirische Analyse junger Unternehmen im Rahmen einer Unternehmensethikforschung*. Wiesbaden: Gabler.
- Wuttke, E., & Seifried, J. (Eds.). (2012). *Learning from errors at school and at work*. Opladen: Barbara Budrich.

Chapter 8

Assessment of Economic Education in Korea's Higher Education



Jinsoo Hahn, Kyungho Jang, and Jongsung Kim

Abstract Economic education has occupied a preeminent and indispensable place in the university curricula in Korea for decades. Considered as one of the most rigorous and prestigious majors, economics departments in Korean universities have invariably attracted applicants of superior grades and qualifications. With this background, this chapter summarizes the current status of economic education in universities in Korea. This chapter also reviews the assessment tools designed to measure economic understanding along with their main characteristics and provide detailed information on the assessment efforts to measure economic literacy at the college level and their main results, more specifically, Test of Understanding of College Economics (TUCE), College Scholastic Ability Test (CSAT), and two nationally accredited qualification tests: the Test of Economic Sense and Thinking (TESAT) and the Test of Economic and Strategic Business Thinking (MK TEST). This chapter calls for a development of a standardized international assessment tool that can measure the economics understanding of college students and adults in hopes to more precisely identify the level of economics understanding of Korean college students by country to country comparison among their similarly situated international peers.

J. Hahn (✉)

Gyeongin National University of Education, Incheon, Korea

e-mail: jshahn@gin.ac.kr

K. Jang

Inha University, Incheon, Korea

e-mail: kjang@inha.ac.kr

J. Kim

Bryant University, Smithfield, RI, USA

e-mail: jkim@bryant.edu

8.1 Introduction

For decades, economic education has occupied a preeminent and indispensable place in the university curricula in Korea. Considered as one of the most rigorous and prestigious majors among social sciences, economics departments in Korean universities invariably attracted applicants of superior grades and qualifications. Korea's rapid economic growth has also contributed to the popularity of economic education and, consequently, the demand for highly trained economists in academics and industries. The inclusion of economics-related subjects in the Korea's civil service exams for high-ranking officials has also prompted many students to study economics, as a major or an elective, in universities. According to a recent education statistic, approximately 77% of universities in Korea, on average from 2011 to 2015, offer economics as a major.

The onset of globalization, furthermore, called for a better economics literacy for students and general public. In response, two nationally accredited civil qualification tests were certified, and they have been utilized as a marker for the college entrance process for some while others used it to obtain college credits, employment, promotion, and/or an evaluation or assessment criterion in firms.

With this background, the authors, in this chapter, summarize the current status of economic education in universities in Korea. Further, we review the assessment tools designed to measure economic understanding along with their main characteristics. This chapter is organized as follows. Section 8.2 surveys the economic education in universities in Korea. Section 8.3 provides detailed information on the assessment efforts to measure economic literacy at the college level and their main results, more specifically, Test of Understanding of College Economics (TUCE), College Scholastic Ability Test (CSAT), and two nationally accredited qualification tests. Section 8.4 concludes.

8.2 Economic Education in Universities

The education system in Korea has been following a 6-3-3-4 platform since 1951. Although high school education is not mandatory, according to an Organisation for Economic Co-operation and Development (OECD) statistic, the proportion of 25–34 years old who completed high school was 98%, the highest among all OECD countries. The corresponding completion rate of 35–44 years old was 94%, one of the highest among the group along with Czech Republic and Slovak Republic (OECD 2011). This is an impressive educational improvement within a generation, in comparison with 43% of 55–64 years old who have the same level of educational attainment. The tertiary graduation rates in Korea reflect this

expansion of access to education. The OECD statistics show that the tertiary graduation rate of 25–34 years old in Korea was 63% in 2011, the highest by far among OECD countries (OECD average is 37% and the European Union 21 (EU21) average is 34% for this age group with Canada following the second with 56%). The high graduation rate in high schools in Korea is partly due to the absence of a failing grade.

The college entrance rate of high school graduates stood at 69.8%, falling slightly below 70% after reaching the peak at 75.4% in 2010. After 2010, the rate has continuously declined but steadily maintaining a 70% plus level until 2014. The high college entrance rate in Korea, somewhat exceedingly high in comparison with other countries, is partly due to its increasingly competitive socioeconomic environment which evolved from the Confucian culture which emphasizes the importance of individuals' educational background.

Approximately 77% (145 out of 188) of Korean universities, averaging 4 years from 2011 to 2015, have economics departments. By academic fields, the proportion of business and economics-related departments takes up 12% in the total number of departments. Within this academic field, the proportion of business administration department is the highest at 6.2%, with international trade/logistics departments at 1.8% and economics department at 1.3% following thereafter. In the case of the colleges of education that are designed to train preservice teachers, the number of departments for social studies education totaled 65, representing only 0.6% of all departments.

Table 8.1 shows the department quotas, the number of applicants, and the newly enrolled. Among the department quotas representing a specific category, business and economics-related departments top the list at 14.6%, with business administration departments at 8.1%, international trade/logistics departments at 1.9%, and economics departments at 1.6%. The social studies education departments account for only 0.4% of the total quota, where economic education occupies mere 513 seats. In terms of college entrance competition rates, business and economics-related departments show 9.3–1, slightly higher than that of engineering and natural science fields at 8.7–1 and 8.6–1, respectively. The competition rate of economics department stands at 10.1–1, higher than that of the business administration department at 9.6–1.

The proportion of male enrollment (54.3%) is greater than that of female enrollment (45.7%). This gender gap, with the predominance of male students, is not as large as the engineering field where the male enrollment surge at 78.4%, but it is higher than that of humanities departments where male students occupy less than 40%. The share of male students in economics and business administration departments is 60.3% and 56.4%, respectively, both of which are substantially higher than the female counterparts.¹

¹ See Hahn (2013) for more information on the trends of economic education in high schools and universities in Korea from 2005 to 2012 and Hahn and Jang (2010) for employment rates by departments.

8.3 Assessments in Higher Education

8.3.1 Test of Understanding of College Economics

There has been very little assessment of the Korean university students' economics understanding. One exception is Jang et al. (2010). This study uses the Test of Understanding of College Economics (TUCE-4) to measure the level of understanding in micro- and macroeconomics for students from ten universities in Korea and compares the findings to those of students in the United States (Walstad and Rebeck 2008) and Japan (Yamaoka 2007).

Table 8.2 shows that the Korean students' level of understanding in micro- and macroeconomics was higher than those of Japan and the United States (USA). In particular, the comparison between Korea and the United States reveals that the gap in economics understanding in microeconomics is larger than that in macroeconomics.

Table 8.2 Overall scores of three countries (Jang et al. 2010)

Content area	Korea			USA			Japan		
	Mean	S.D	N	Mean	S.D	N	Mean	S.D	N
Microeconomics	58.2	15.8	992	42.0	15.6	3,859	41.7	14.9	448
Macroeconomics	53.1	20.5	1,163	46.9	17.6	3,495	38.6	19.9	408

Next, the analysis of content categories classified by Walstad and Rebeck (2008) shows that Korean university students' mean scores are higher than those of Japanese and the US university students in all categories (Table 8.3). The largest gap was found in the "basic problem" category in which Korean university students' performance is the highest at 70.6 points and the US university students' performance is the lowest at 39.5 points. In macroeconomics, Korean students outscored everyone with 62.3 points in the "measuring aggregate performance" category, while the US students outperformed everyone in the "aggregate supply and demand" category with 54.5 points. The largest gap between the US and Korean students was in "international (macro)" economics, where Korean students outperformed the US students by 12.1 points.

Table 8.3 Mean scores for content specifications (Jang et al. 2010)

Content area	Contents	Korea	USA	Japan
Microeconomics	Basic problem	70.6	39.5	57.2
	Markets and prices	64.0	40.6	46.6
	Theories of firm	50.0	46.1	34.9
	Factor markets	58.2	41.3	41.3
	Micro role of government	53.7	40.6	39.9
	International (micro)	68.6	40.3	42.6

(continued)

Table 8.3 (continued)

Content area	Contents	Korea	USA	Japan
Macroeconomics	Measuring aggregate performance	62.3	53.0	43.6
	Aggregate supply and demand	55.1	54.5	42.0
	Money and financial markets	49.3	45.8	42.1
	Policy debates and applications	53.0	45.4	39.6
	Macro role of government	40.5	35.0	23.3
	International (macro)	54.4	42.3	32.7

The comparison of micro- and macroeconomics scores by the cognitive categories classification by Walstad and Rebeck (2008) shows that Korean students outperform US and Japanese students across all categories in microeconomics (Table 8.4). In macroeconomics, on the other hand, Korean students perform better than US and Japanese students in two out of three categories. In microeconomics, the largest gap was found in the “explicit application” category with Korean students performing at the top with 64.9 points and the US students at the bottom with 40.8 points. In macroeconomics, both Korean and US students best perform in the “explicit application” category within each country’s sample population. Korean students, however, outperform the US students with a 7.6-point margin when the resulting numbers are compared in this category. Korean students also perform better than the US students in the “implicit application” category by a 5.9-point margin. In the “recognition and understanding” category, on the other hand, the US students outperform the Korean students by a small margin.

Table 8.4 Mean scores for cognitive specifications (Jang et al. 2010)

Content area	Cognitive categories	Korea	USA	Japan
Microeconomics	Recognition and understanding	46.9	43.8	31.3
	Explicit application	64.9	40.8	46.7
	Implicit application	53.3	43.3	39.7
Macroeconomics	Recognition and understanding	46.8	47.0	39.3
	Explicit application	56.4	48.8	38.4
	Implicit application	51.1	45.2	38.5

To explain the Korean students’ better performance, the sample group’s makeup must be considered. One explanation may be that more than 50% of the Korean sample students are from the social studies education department which has a higher bar when it comes to college entrance test scores for admittance (Jang et al. 2010). The use of relative grading is suggested to compare the abilities of economics understanding among countries (Jang et al. 2010). The relative grading can be obtained by comparing grades by contents or by cognitive categories to the total score.

The comparison of relative grades by content categories shows that Korean students perform better than the US students in categories such as “basic problem,” “market and prices,” and “international (micro)” in microeconomics. But Korean

students show a relatively lower level of understanding than the US students in “theories of firm” and “micro role of government” categories. In macroeconomics, the Korean students display a relatively higher level of understanding than the US students in “measuring aggregate performance,” “policy debates and application,” and “international (macro)” categories. But in both “aggregate supply and demand” and “money and financial markets” categories, the Korean students show relatively lower understanding than the US students. The comparison of relative scores by cognitive categories shows that the Korean students perform relatively better than the US students in the “explicit application” category both in micro- and macroeconomics. But Korean students show relatively lower understanding in the “recognition and understanding” category.

8.3.2 College Scholastic Ability Test

8.3.2.1 Format of the Test

Administered once a year in early November and required for almost all college applicants, the College Scholastic Ability Test (CSAT) is designed to assess higher-order thinking skills and provide reliable information for college admission. Although CSAT is not targeted to university students, it is worth reviewing the test as it measures Korean students' ability for college education.

Notwithstanding the fact that the American College Testing (ACT) used in the United States has no economics subject section, the conceptual framework is pretty much the same as CSAT. Both tests share a rationale that can be effective not only in assessing the students' level of understanding in economics but also predicting students' potential achievement and study success in tertiary education (Zlatkin-Troitschanskaia et al. 2015).

There are both pros and cons of using the CSAT economics subject test to assess the high school seniors' (including repeaters) level of understanding in economics. Since the test results substantially affect student's college admissions, students strive to do their best by paying attention to meticulous details. Thus, the upside of CSAT is that little or no apathetic attitude or blanks are indicated in the answer sheet. Due to a large number of test takers, the degree of reliability in the results is greatly enhanced.

On the other hand, a self-selection bias problem may occur since the students who took economics courses generally choose to take the CSAT economics test. This bias may exaggerate the level of economic literacy of the average Korean high school graduates. Therefore, the CSAT economics test results should be interpreted as representing the economic literacy of high school students who took economics courses in school rather than all high school students.

The objectives of the CSAT are not limited to measuring students' memorization and recall skills but also assess the analytic thinking skills such as abilities to apply knowledge and solve problems and inferential reasoning. To achieve these assessment goals, CSAT questions consist of a three-part structural format (see the exam-

ple question in Sect. 8.3.2.2). First, a heading that includes directions for problem-solving or a description of the question precedes the test problems. All headings are in question forms.

Second, materials are provided for exploring the problems. The types and forms of materials are diverse. In case of economics, literature texts, news articles, data tables, various types of chart, and cartoon/comics are used. Students are required to choose the correct answer by understanding, interpreting, analyzing, and inferring from the given materials with appropriate utilization of economic concepts and principles.

Third, answer choices are presented. In general, five answer choices are given. If there is a need to make questions harder to answer, students are asked to choose a set of correct choices among four choices presented in a box. Since students perceive this format more difficult, such questions are allowed up to 30% of the total test. Since the CSAT questions are created to be investigative in nature, a question on average is allotted 90 s to reach a correct answer (20 questions in 30 min). This is about twice as long as the US History Subject Test, a 60-min test consisting of 90 multiple-choice questions, with an allotted time of 40 s per question.

8.3.2.2 The Frequency and Percentage of Correct Responses by Areas

The CSAT does not follow the question bank format. Every year, specialists consisting professors and teachers make new questions for the test. An economics test, consisting of 20 questions, includes contents from microeconomics, macroeconomics, and international economics. Based on the difficulty of questions and the importance of curriculum content, there are 10 questions, each weighing 2 points with another 10 questions, each weighing 3 points, totaling 50 points. There is no penalty for incorrect answers.

Table 8.5 shows the frequency and percentage of correct responses to the economics test questions from 2004 to 2015 by content areas. Out of the total 240 questions, the ratio of microeconomics questions is more than half at 56.7%. The remaining questions are allocated to macroeconomics and international economics, at 25.8% and 17.5%, respectively.² These patterns reflect the state of present economics curriculum in Korea for high school students, where out of six chapters in high school economics textbooks, four chapters are devoted to microeconomics and only two chapters are devoted to the study of macroeconomics and international economics with one chapter allocated to each.³

²The items (in Korean) can be obtained from <http://www.suneung.re.kr/boardCnts/list.do?boardID=1500234&m=0403&s=suneung>.

³The National Assessment of Education Progress (NAEP) test in the United States recommends the distribution of 45% for microeconomics, 40% for macroeconomics, and 15% for international economics (Buckles and Walstad 2008).

Table 8.5 Frequency and percentage of correct responses by content areas

Content area	Frequency (%)	Percent correct	<i>F</i> -statistics (<i>P</i> value)
Microeconomics	136 (56.7)	58	10.205 (0.000)
Macroeconomics	62 (25.8)	54	
International economics	42 (17.5)	44	
Total	240 (100.0)	54	

Note: The numbers in the parentheses are the percentage shares of content area test questions appearing in all the tested questions. Percent correct is the mean percentage of correct responses

One method of analyzing the CSAT assessment results is to investigate the pattern by the percentage of correct responses. Table 8.5 shows the mean percentage of correct responses by content areas.⁴ The percentage of correct responses, which can estimate the difficulty of test questions, theoretically lies between 0 and 100, but in practice, most questions lie between 20 and 80. The mean percentage of correct responses for all CSAT economics questions for 12 years is 54, which indicates that the test takers perform better in microeconomics. On the other hand, the mean percentage of correct responses for international economics stands only at 44, indicating that test questions related to trade and exchange rates are difficult. The mean percentage of correct responses between microeconomics at 58 and international economics shows a gap as much as 14%. The authors fail to reject the hypothesis that there are differentials in the mean percentage of correct responses by content areas at 1% significance level ($F = 10.205$, $P = 0.000$).

Since the CSAT faithfully reflects the national curriculum stipulated by the Ministry of Education, Korea's economics curriculum can be understood by the frequency of test questions by content standards appearing on the test.⁵ In the United States, the Center for Excellence in Education (CEE) (2010) promulgates 20 standards for economic concepts and principles, which are reflected in content standards 1–20 in Table 8.6.

The majority of learning elements included in the economics curriculum in Korea overlaps with these 20 standards, but there are also learning elements found only in the country's economics curriculum, which are classified as standards 21–24 (elasticity, aggregate supply and demand, balance of payment, and personal finance). Table 8.6 reports the CSAT questions by 24 standards. With the authors' debated discussions and further considerations, for the purpose of this study, questions including more than one standard have been classified as the core assessment factors standard.

⁴The percentage of correct responses of test questions is not publicly released. The percentages presented in Table 8.5 are drawn from the consensus among authors and multiple institutions which collected a variety of relevant information.

⁵The Ministry of Education, Science and Technology was renamed as the Ministry of Education in 2015. See Ministry of Education, Science and Technology (2012) and Ministry of Education (2015) for the national economics curriculum.

Table 8.6 Frequency and percentage of correct responses by content standards

Content standard	Selected key concepts	Frequency (%)	Percent correct
1. Scarcity	Scarcity, choice, opportunity cost, factors of production, economic activity	7 (2.9)	64
2. Decision-making	Marginal benefit and cost, utility maximization, profit maximization, trade-off	15 (6.3)	62
3. Allocation	Three basic economic questions, economic system, market economy, planned economy	9 (3.8)	71
4. Incentives	Incentives, people's response	2 (0.8)	86
5. Trade	Gains from trade, exports, imports, trade barriers	6 (2.5)	56
6. Specialization	Specialization, division of labor, comparative advantage, interdependence	12 (5.0)	40
7. Markets and prices	Market, price, quantity demanded, quantity supplied, exchange rates	5 (2.1)	58
8. Role of prices	Role of prices, changes in supply and demand, price control	43 (17.9)	51
9. Competition and market structure	Competition, monopoly, oligopoly, collusion	3 (1.3)	69
10. Institutions	Banks, saving, property rights, household, labor unions, corporation	3 (1.3)	57
11. Money and inflation	Roles of money, money supply, price, inflation	0 (0.0)	–
12. Interest rates	Interest rate, nominal vs. real, present value	3 (1.3)	38
13. Income	Income, labor, human capital, wages, productivity, distribution	8 (3.3)	41
14. Entrepreneurship	Entrepreneurs, innovation, risk, profit, invention	1 (0.4)	91
15. Economic growth	Productivity, standard of living, investment, human capital, economic growth, GDP	17 (7.1)	51
16. Role of government and market failure	Roles of government, public goods, externalities, redistribution, regulations	29 (12.1)	57
17. Government failure	Government failure, political leaders, interest groups	0 (0.0)	–
18. Economic fluctuations	Nominal GDP, real GDP, business cycle, recession, circular flow diagram	17 (7.1)	58
19. Unemployment and inflation	Unemployment rate, labor force, inflation costs, purchasing power	18 (7.5)	56
20. Fiscal and monetary policy	Budget, fiscal policy, budget deficit, debt, monetary policy	9 (3.8)	55
21. Elasticity	Price elasticity of demand, elastic, inelastic, revenue	16 (6.7)	46
22. Aggregate supply and demand	Aggregate supply, aggregate demand	3 (1.3)	48
23. Balance of payment	Current account, goods account, services account	9 (3.8)	45
24. Personal finance	Financial product, money management, liquidity, risk, return, life cycle, credit	5 (2.1)	75
Total	–	240 (100)	54

Note: Content standards 1–20 are from CEE (2010), and 21–24 are prepared by the authors in light of the Korea's economics curriculum

The area from which the most CSAT economics questions are asked is the “role of prices,” comprising approximately 18% of total questions. This is both logical and consequential in that the principles of market price determination and changes are the predominant components in economics. The Korean economics curriculum also treats this area as one of the core contents.

In order to assess the students’ understanding of changes in demand and supply on the equilibrium price, the effect of price changes on individuals’ choices, and the consequences of price controls on market behaviors and their respective impacts thereof, CSAT has utilized a variety of examples of goods and services. For instance, the following question asks about the exchange rate determination in the foreign currency market. Gleaning from the conversation between the characters in the cartoon, students are asked to choose the correct answer by making an inference about the changes in the foreign currency market (Fig. 8.1).

From the reading of the conversation between the characters below looking at a newspaper article, what would be the most appropriate title of the article?



1. The number of students going abroad for study rises.
2. The foreigners’ purchase of domestic bonds rises.
3. Domestic firms’ foreign investment declines.
- 4. Due to economic downturns in international market, export declines.^a**
5. Due to the falling price of raw materials in international market, import rises.

Fig. 8.1 Assessment 1: The exchange rate determination in the foreign currency market

^aThe choice in bold letters is the correct answer

The market failure and the role of government are the second most frequently asked area. Twelve percent of total questions assess the students' understanding on externality, the characteristics of public good, the market failure, and the role of government. Other frequently asked areas are the cost-benefit analysis and rational choice, the relationship between gross domestic product (GDP) and gross national product (GNP), the distinction between nominal and real national income, and the composition of labor force and unemployment rates.

The price elasticity of demand is regarded as an important concept in Korea's economics curriculum. Accordingly, economics teachers emphasize the important fact that the impact of price change on total revenue varies depending on the size of price elasticity of demand, and to assist students to better understand this principle, they teach by providing specific examples. Therefore, the frequency of questions on elasticity is relatively high (6.7% of total questions). The following question, for example, assesses a student's ability to understand how the price elasticity of demand affects the shape of a demand curve for liquid crystal display television (LCD TV) and compare the impact arising out of the decrease in supply on the market equilibrium under two types of elasticity (Fig. 8.2).

Which choices in the <Box> are correct inferences that can be derived from the market prospects of LCD TV made by security companies X and Y in the following situation?

While it is expected that the price of LCD panel, a key component of LCD TV, will rise by 10%, the different scenarios projected by security companies X and Y on the LCD TV market make investors confused. The company X presumes that the price elasticity of demand is elastic in the LCD TV market, but company Y presumes that it is inelastic.

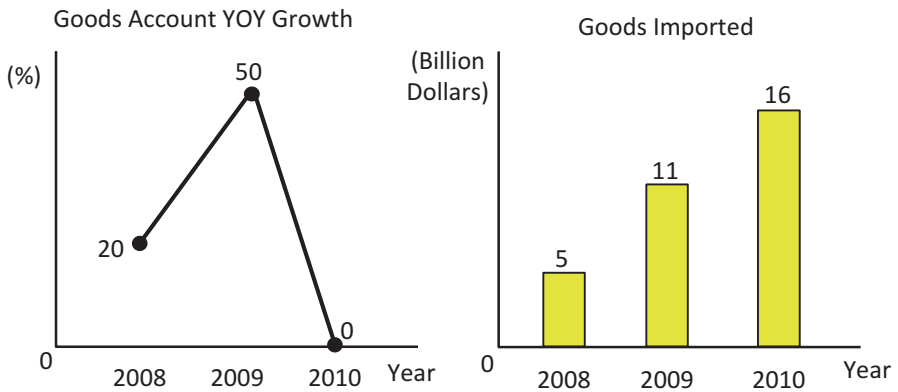
< Box >

- A. In comparison with company Y, the company X expects that the price will change more.
- B. In comparison with company X, the company Y expects that the quantity demanded will change less.
- C. Both companies X and Y expect that the equilibrium price will rise and the equilibrium quantity demanded will decline.
- D. Company X expects that sales revenue will rise, but the company Y expects that sales revenue will decline.

A, B
 A, C
 B, C
 B, D
 C, D

Fig. 8.2 Assessment 2: The price elasticity of demand

One of the features of CSAT is that questions measuring the ability to interpret data and graphs are frequently asked. The intention of this practice is to assess a student's ability to understand and interpret the statistical data that are presented in diverse forms such as graphs, charts, and tables. This is because one of the underlying goals of economic education is to equip the general public who did not major in economics with the ability to understand economics-related news and articles from the media outlets which contain a diverse form of statistics. For example, the following question, by using the data on the growth rate, assesses the test taker's ability to figure out the balance of goods account and accurately calculate the amount of exports (Fig. 8.3).



< Box >

- A. The balance of goods account in 2008 is 7 billion dollars.
- B. Since 2008, the amount of goods exported has consistently increased.
- C. The YOY amount of change in goods exported is the largest in 2010.
- D. There is no difference in exports and imports in years 2009 and 2010.

① A, B ② A, C ③ B, C ④ **B, D** ⑤ C, D

Fig. 8.3 Assessment 3: The balance of goods account

The following two charts show the growth rate of goods balance account (YOY) and the dollar amount of goods imported. Which choices in the <Box> are correct? (The goods balance in 2007 is five billion dollars).

Another characteristic in CSAT is the inclusion of new questions about personal finance. The financial crisis of 2007 heightened the public awareness and the importance of personal financial literacy. Consequently, fundamental components of personal finance were added to the economics curriculum when it was revised in 2009. As a result, out of six chapters in high school economics textbooks, one chapter has been assigned to personal finance.

After the inclusion of personal finance in the high school curriculum, approximately two questions have been asked each year to assess students' financial literacy. It is noteworthy to mention that in the Korean economics curriculum, the questions about nominal and real interest rates fall in the category of personal finance, but they are classified as interest rates (standard 12) in the CEE classifications in Table 8.6. Therefore, if 3 questions classified as standard 12 are added to standard 24, 3.3% (8 questions) of total questions incidentally fall into the personal finance category. In light of the fact that the inclusion of personal finance questions began in 2012 and the prominence of personal financial responsibility gained since then, the share of such questions will only increase in the future.

On the other hand, there currently are no CSAT questions that test the students about roles of money, inflation, and government failure. Although Korea's economics curriculum explicitly includes roles of money, the relationship between money supply and price level, and inflation, these concepts are generally regarded as areas merely requiring memorization of concepts and definitions, consequently, less attractive for CSAT questions. In addition, questions about incentives and entrepreneurship have been rarely included.

The pattern arising from the percentage of correct responses shows that questions about incentive, entrepreneurship, and personal finance are relatively higher in percentage of correct responses, while questions about the terms of trade for mutually beneficial trade, Lorenz curve and income distribution, and balance of international trade are relatively lower.

However, we need to interpret this pattern with caution since the number of questions falling into these categories is relatively small. Focusing narrowly on the standards from which relatively many questions (ten or more) are asked, the percentage of correct responses is higher for questions related to rational decision-making, cost-benefit analysis, market failure and the role of government, and the circular flow of economic activities. Lower correct responses were registered for questions related to the comparative advantage and the price elasticity of demand.

8.3.2.3 Overall Difficulty

Figure 8.4 presents two data that show the degree of difficulty of CSAT. The line above indicates the average percentage of correct responses by year. The trend reveals that since 2004, the average percentage of correct responses had consistently declined until 2010 as the questions have become more difficult, but since 2013 the questions have become easier. Considering that the goal of the CSAT's level of difficulty is to ensure approximately 50–70% of correct responses, one can see that economics test questions have been excessively difficult during the 6-year period from 2006 to 2012. In particular, the economics tests were most difficult from 2008

to 2010. It appears that the level of difficulty in economics test was restored to an appropriate level since 2013.

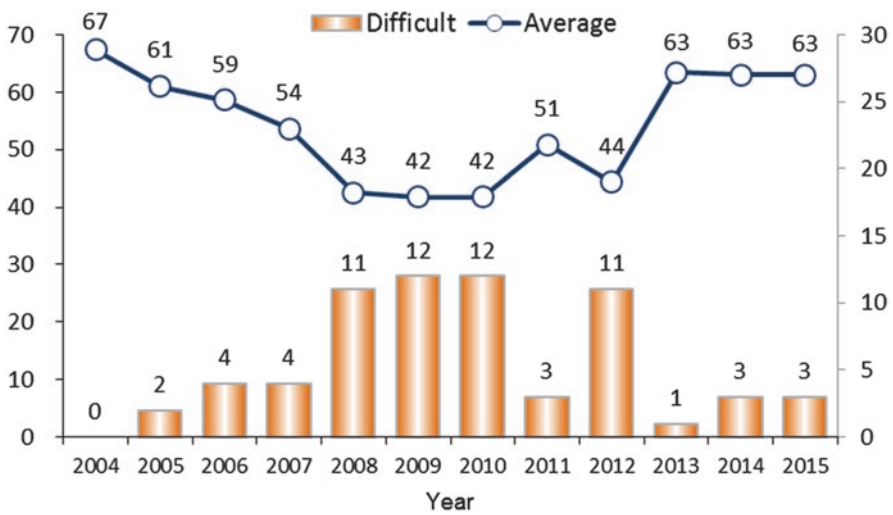


Fig. 8.4 Degrees of CSAT's difficulty, 2004–2015

Note: The line graph indicates the average percentage of correct responses by the CSAT test takers (left axis), and the bar chart shows the number of difficult questions with 40% or less of correct answers (right axis)

The level of difficulty of the CSAT economics test can be confirmed by the frequency of questions regarded as difficult. According to the standards promulgated in US National Assessment of Educational Progress (NAEP) exam classification, difficult questions have less than 40% of correct answers, while medium and easy questions have 40–60% and more than 60% of correct answers, respectively. The ideal exam would consist of 10–12 questions of medium difficulty and 4–5 easy and difficult questions each.

But as the bar graph in Fig. 8.4 shows, the number of difficult questions increased to 11–12 in 2008–2010 and again in 2012. The economics test questions became easier since 2013. In contrast to the period of 2008–2010, the number of difficult questions greatly decreased, and the number of easy questions increased. In this regard, from the perspective of the level of difficulty, the CSAT economics test has not been satisfactory in general.

This unsatisfactory result regarding test difficulty is attributed to its fundamental systemic procedure with limits the question creation process. The CSAT, administered once a year, does not adopt the method of question bank system due to security reasons. Instead, the CSAT questions are created in a closed form by professors and teachers who gather and stay in a high security area for about 30 days immediately prior to the test. Since a small number of specialists are burdened with a time constraint to quickly develop test questions, gauging the right degree of difficulty in test questions and consistently maintaining the same level of intensity in test questions every year become increasingly difficult.

8.3.2.4 Standardized Scores by Gender

Considering the degree of difficulty which varies across test subjects, CSAT releases not the original but the standardized test results. Although individual test scores are not released, the gender difference in economics understanding can be analyzed from the frequency distribution of standardized scores by gender.

Figure 8.5 shows the relative cumulative frequency of standardized scores by gender. Before examining the relative cumulative frequency, it is important to focus on two facts. First, the share of male students in total economics test takers has been consistently rising. In 2006, the share of male student was 53.3%, but in 2015, the share exceeded two thirds. Second, during the period when economics tests were difficult, the standardized score of students who received full marks rose substantially. For example, in 2009, the standardized score of students who scored 100% in the test was quite high at 81, but in 2015 when the test questions were easy, the standardized score was only 69.

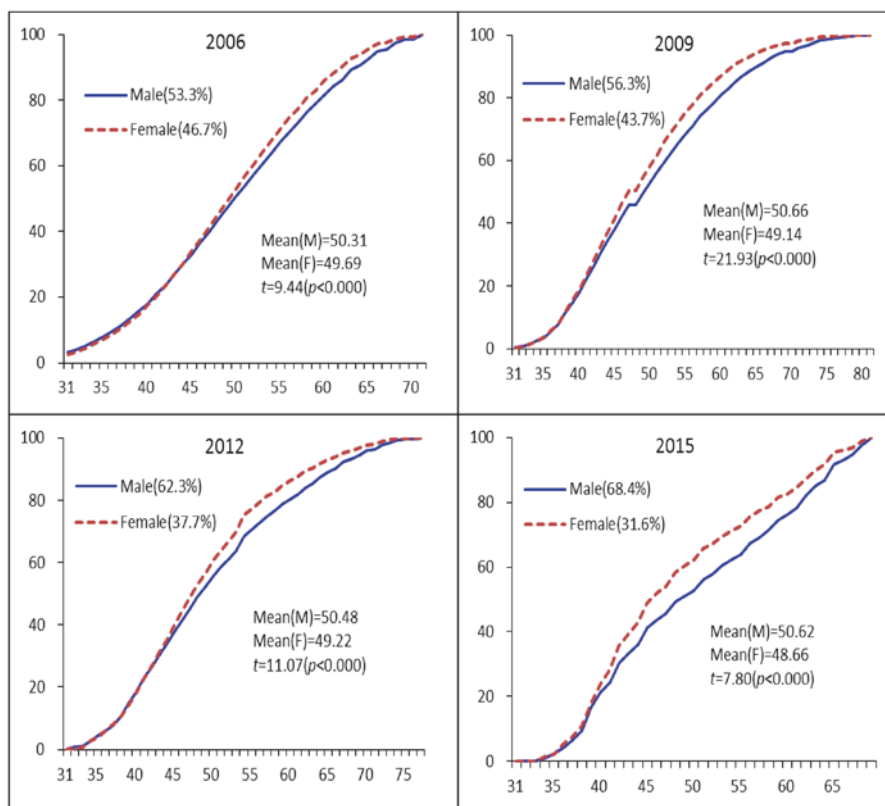


Fig. 8.5 Relative cumulative frequency by gender

Since the relative cumulative frequency by gender shows a similar pattern from year to year, the results for only the 4-year periods are presented in Fig. 8.5. Several interesting patterns were found. First, in the lower range of economics test scores, the number of male students is slightly higher than that of female students, but the differences are not statistically significant. Second, in the middle range which covers the largest segment of the entire spectrum of scores, the frequency of female students is higher than that of male students. Third, in the upper range of scores, the frequency of male students is higher than that of female students.

The implication from Fig. 8.5 is that, during this period, the level of understanding in economics of male students was higher than that of female students, and the gender gap has widened. For example, in 2006, the standardized scores were 50.31 for male students and 49.69 for female students, but in 2015 the male students' score rose to 50.62, while the female students' score declined to 48.66. Test results also found that the level of male high school students' level of economics understanding was higher than that of female high school students.⁶

8.4 Nationally Accredited Civil Qualification Tests Introduction on Background

Due to curriculum changes, among other reasons, the status of economic education in Korean high schools has continued to fall in the last decade or so. To remedy this decline, private institutes as well as the Korean government have explored a variety of methods to broaden the horizons of economic education to both students and the general public. In the process, arguments were put forth in support for the need to introduce a nationally accredited test. According to the survey results conducted by Moon et al. (2010), 66.9% of the 320 surveyed managers who were engaging in diverse industries answered positively to the introduction of a nationally accredited test. About a half (44.4%) of the respondents were willing to apply the outcome of the nationally accredited tests to their own companies. For those who were willing to use the test results in the companies of their employment, 70.9% of respondents indicated the overriding reason for taking the test was for education or assessment purposes and 59.1% of respondents preferred grade level to raw scores for their evaluation method.

Against this backdrop, two major Korean business newspaper companies launched nationally accredited civil qualification tests. Although anyone can take these two tests regardless of age, the majority of test takers are university students.

⁶Most studies on male-female differences in economic literacy find that males performed better than females. For a comprehensive survey on the gender gap, see Siegfried (1979). Recently, Walstad and Buckles (2008) and Walstad (2013) also confirm that male students significantly out-scored female students in the 2006 NAEP economics test. Similar gender effects were also founded in a cross-national study by Brückner et al. (2015).

In this regard, it is worth reviewing these tests to discuss issues regarding the assessment of economics understanding in Korea's higher education environment.

In November 2010, *The Korea Economic Daily* launched its own test, called TESAT (Test of Economic Sense and Thinking). A month later in December 2010, the *Maeil Business News Korea* also launched its own test, MK TEST (Test of Economic and Strategic Business Thinking). The launch of nationally accredited civil qualification tests appears to resemble the Japanese case in which a nationally accredited civil qualification test was introduced by the *Nikkei* newspaper.

The TESAT has been administered six times a year since 2015, increased from four times. The MK TEST has been administered eight times a year since 2015, increased from four times. For each test, approximately 2000–4000 individuals take the test. The test grades are utilized in processing employment, promotion, and assessment in finance-related public companies as well as financial companies and major private firms. The test results are also used to acquire 14–20 college credits through academic credit bank system for students pursuing a bachelor's degree.

8.4.1 TESAT

The TESAT consists of total 80 questions with maximum 300 possible points (Table 8.7). There are three categories in TESAT: 30 questions pertain to economic theory, with another 30 questions on current economic affairs, and with the remaining 20 questions on inference and decision-making. Each section carries 100 points. The length of the test is 100 min so that each question is allowed no more than 1 min and 15 s on average to answer. Accordingly, TESAT, unlike CSAT, includes a larger share of questions about basic knowledge rather than questions that require deeper analysis and investigation.

Table 8.7 Composition of TESAT questions (<http://www.tesat.or.kr>)

Content area		Abilities (points)			Total	
		Knowledge understanding (3)	Application (4)	Analysis, inference, comprehensive decision (5)	No.	Points
Economic theory	General basics	20	10	–	30	100
	Microeconomics					
	Macroeconomics					
	Finance					
	International economics					
Current economic affairs	Policy/statistics	20	10	–	30	100
	General knowledge/terminology					
	Business					

(continued)

Table 8.7 (continued)

Content area		Abilities (points)				Total	
		Knowledge understanding (3)	Application (4)	Analysis, inference, comprehensive decision (5)			
					No.	Points	
Inference and decision-making	Data interpretation	–	–	20	20	100	
	Issue analysis						
	Decision-making						
Total	No. of questions	40	20	20	80	–	
	Points	120	80	100	–	300	

The TESAT notifies to its test takers the assessment results in absolute (raw score) in form of grades and relative (scaled score) in form of percentile ranks. For score grades, there are six grade categories and no-grade (unqualified/failed) category. In the grade category S is the highest that requires over 90 out of 100 converted points (100 converted points are equivalent to the perfect raw score of 300 points). Next come the grade 1 (80–90 points), grade 2 (70–80 points), grade 3 (60–70 points), grade 4 (50–60 points), and grade 5 (40–50 points). No-grade category is assigned if the score is lower than 40 points, which is deemed as unqualified. The nationally accredited qualification status is awarded to grade 3 and better (over 60 converted points). The qualification remains valid up to 2 years from the date the scores are announced.

Table 8.8 lists the characteristics of test takers and their grades (scores) for TESAT from its 18th test in 2013 to 34th test in 2016. For test takers' characteristics, the proportion of students attending university or graduate school is the largest at 46.86%, followed by high school students at 25.67%, and job applicants at 11.95%. The proportion of company workers is merely 8.37%. This implies that TESAT is more utilized by students to obtain a nationally accredited qualification to gain admission to universities or those seeking employment rather than by workers seeking promotion and personal evaluation in private sector companies. This also can be confirmed by the large proportion of test takers (40.12%) who are upper classmen in their universities. By majors, the proportion of test takers who major in business or economics is the highest at 39.22%, followed by humanity/social science majors. On the other hand, the proportions of engineering majors, natural science majors, and arts and physical education majors are much lower.

Table 8.8 TESAT applicant's scores (2013–2016, average) (<http://www.tesat.or.kr>)

Category	Sub-category	Share (%)	Economic theory	Current affairs	Inference and decision	Total score
Grade	S (90–100)	2.44	93.62	91.14	93.41	92.72
	1 (80–90)	11.44	85.78	81.77	84.66	84.07
	2 (70–80)	19.28	76.27	72.37	75.03	74.56
	3 (60–70)	21.74	66.40	62.74	65.55	64.90
	4 (50–60)	19.02	56.32	53.49	54.85	54.89
	5 (40–50)	14.08	46.53	44.14	44.86	45.17
	Unqualified (0–40)	12.00	34.88	33.06	31.97	33.30
Occupation	College and graduate students	46.86	63.83	59.94	62.18	61.98
	Company workers	8.37	52.90	56.43	53.44	54.26
	Self-employed	0.35	52.59	57.08	50.03	53.23
	Job applicants	11.95	65.08	63.98	63.85	64.30
	Military personnel	2.42	71.87	66.99	71.25	70.04
	Others	4.20	57.41	55.71	55.54	56.22
	High school students	25.27	65.14	60.02	63.59	62.92
Major	N/A	0.59	60.63	59.19	57.76	59.19
	Business/economics	39.22	65.02	62.43	63.60	63.68
	Humanity/social science	19.18	63.55	60.18	61.89	61.87
	Natural science	3.77	59.79	57.56	59.36	58.90
	Engineering	8.02	52.73	53.04	52.64	52.81
	Arts and physical education	0.86	47.24	48.07	46.25	47.19
	Others	3.56	57.12	55.14	55.29	55.85
Education	N/A	25.38	65.05	59.99	63.41	62.82
	University (1 and 2 years)	8.93	60.29	54.93	58.52	57.91
	University (3 and 4 years)	40.12	64.70	61.49	63.24	63.14
	University graduates	19.14	61.25	61.99	60.95	61.39
	Graduate school graduates	1.31	61.30	63.03	61.09	61.81
	High school graduates	1.84	52.15	50.59	49.84	50.86
	Others	6.15	59.45	56.09	57.49	57.68
Total	N/A	22.50	65.16	60.18	63.55	62.96
		100.00	62.96	59.98	61.50	61.48

Note: Scores are converted into 100 from total 300 points

With respect to scores, the average score is 61.48 points. The score for economic theory is somewhat higher at 62.96 points compared to other areas. The score gap, however, in comparison with other areas, is not significant. With respect to grades, those who scored over 90 out of 100 points represent 2.44%. The proportion of test takers who received grade 1 constitutes 11.44%, while grade 2 represents 19.28% of the tested group. While the proportion of test takers who received grade 3 and better is 54.90%, 12.0% of test takers scored less than 40 points, failing to attain any grade.

With respect to occupation or student status of the test takers, job applicants performed best with the score of 64.30 points, followed by high school students, university students, and company workers. This finding substantiates the belief that TESAT is used more for job application and college admission purposes than for promotion-related personal assessment in companies. In particular, the high performance of high school students implies that top-ranked high school students use the nationally accredited civil qualification tests to strengthen their portfolios for college admissions.

By majors, test takers from business/economics-related majors performed the best at 63.68 points, followed by humanity/social science majors and natural science majors. The finding of high performance (62.82 points) for test takers who did not specify their majors implies that they were high school students. By educational attainments, test takers who are junior or seniors in universities performed the best at 63.14 points, followed by graduate degree holders and bachelor's degree holders.

8.4.2 MK TEST

The MK TEST consists of 80 questions, same as the TESAT. But unlike the TESAT, the total score of the MK TEST is 1000 points (Table 8.9). The test is divided into economics and business with 40 questions in each section with 500 points for a combined total of 1000 points. Within each area, there are 15 questions on knowledge, 15 questions on reasoning ability, and 10 questions on current affairs. The questions on the reasoning ability are weighed more heavily than those on knowledge or current affairs. The MK TEST is 90 min long, 10 min shorter than the TESAT. Accordingly, each question in the MK TEST is assigned on average no more than 1 min 8 s.

Table 8.9 Composition of MK TEST questions (<http://exam.mk.co.kr>)

Content area		Abilities (points)			Total	
		Knowledge (10)	Reasoning (17)	Current affairs (10)	No. of questions	Points
Economics		15	15	10	40	500
Business		15	15	10	40	500
Total	No. of questions	30	30	20	80	–
	Points	300	500	200	–	1000

Note: The points in the table were obtained by dividing the subtotal points by the number of questions in each field

The MK TEST not only notifies the measure of absolute assessment (grade) but also the measure of relative assessment (scaled score) in percentile ranks. There are four grades: “excellent” is awarded to test takers who received over 80 out of 100 points, “good” for 60–80 points, “fair” for 40–60 points, and “unsatisfactory” for less than 40 points. The nationally accredited qualification status is awarded to those who received 60 or more points (excellent or good grade), and the qualification remains effective for 2 years from the award date. The score certificate of the MK TEST includes information on the scores and percentile ranks for each area (knowledge, reasoning ability, and understanding of current affairs) in economics and business.

Table 8.10 shows the test performance, based on the 16th and 17th test results of economic section, sorted by test takers’ gender, occupation, and majors. Males performed better than females. In the 16th MK TEST, males’ grade was 63.2 points, outperforming the females by 6.3 points. In the 17th MK TEST, male test takers also outperformed their female counterparts by 6.0 points.

Table 8.10 MK TEST takers’ performance (16th and 17th tests) (Maeil Business News Korea, internal data)

Category		16th			17th		
		Male	Female	Total	Male	Female	Total
Occupation	Middle school	55.0	–	55.0	60.0	–	60.0
	High school	67.6	58.6	64.7	67.0	60.7	65.2
	University (1–2)	60.3	52.5	57.4	51.3	42.2	46.9
	University (3–4)	65.6	59.2	63.0	60.4	54.2	57.6
	Job applicant	63.8	61.6	62.9	62.7	58.5	60.6
	Salaried man	54.5	49.8	52.8	48.3	43.5	46.5
	Government worker	53.3	68.8	59.5	57.5	40.0	47.0
	Professional	60.0	43.6	49.5	41.3	44.7	43.2
	Self-employed	52.8	32.5	50.9	43.1	34.2	41.3
	Soldier	62.6	–	62.6	60.0	–	60.0
	Others	62.4	51.6	58.2	57.0	47.1	52.8
No response	63.5	54.2	60.0	62.1	54.7	58.9	
Major	Business, economics	65.1	57.9	62.3	61.7	55.5	59.1
	Humanity, social science	62.8	56.6	60.2	59.3	52.6	55.8
	Education	70.0	61.3	64.7	69.4	53.7	57.4
	Natural science	60.2	62.1	61.0	49.9	48.8	49.2
	Engineering	54.5	54.7	54.5	51.1	49.6	50.8
	Arts and physical education	62.5	54.2	57.5	50.7	48.9	49.9
	Others	60.2	50.0	55.8	52.0	43.8	48.4
	No response	64.5	54.7	61.2	61.6	54.7	58.8
Total	63.2	56.9	60.8	59.5	53.5	56.9	

Note: Since the information on the points of each question was not released, the authors apply 2.5 points equally to all 40 questions in economics field to calculate scores

By current occupation status, like the TESAT, the performances of high school students, university junior/seniors, and job applicants are found to be higher than that of employed workers seeking to use the test for internal promotion within corporations. To explain these groups' high performance, Song et al. (2015) point out that they tend to take the test to reach certain target or grade level for college admission or job application purpose, by repeatedly solving previous test questions. Lastly, business, economics, and education concentration students outperform students of other majors.

8.5 Conclusion

While the curriculum changes, among other reasons, have contributed to reducing the extent of economic education in Korean high schools, economic education in Korean universities has become more comprehensive in the general assessment levels. Korean university students, in outperforming their Japanese and US peers in both micro- and macroeconomics in TUCE, may reflect the quality of economic education in Korean universities.

In this chapter, we investigated the economic literacy of Korean high school students by using the results of CSAT in addition to TUCE. The CSAT is a country-specific test compared to TUCE, but it offers some incentives in that it is assessed for a large number of students every year. In particular, the standardized scores of the CSAT are released separately by gender. The data show that the level of economics understanding of Korean male high school students is higher than that of female high school students, and the gender gap has widened recently.

We also presented the test results for TESAT and MK TEST, designed to assess economic and business literacy for students and adults with various educational backgrounds. These two nationally accredited civil qualification tests were introduced to broaden the horizons of economic education and have been successfully implemented as certified tests. These tests are also important for testing economic literacy since they have been given more frequently, as many as eight times a year. Through these tests, a little over half of all applicants earned the nationally accredited qualification status, granted to applicants who score over 60 out of 100 points. The job applicants were found to have higher level of economics understanding compared to company workers and the self-employed. The economics understanding of individuals who major in natural sciences, engineering, or arts and physical education was below the national certification level on average, and the economics understanding levels of the juniors and seniors were found to be higher than that of freshmen and sophomores. Accordingly, more efforts should be expended to strengthen the economic education for those whose economic understanding level is relatively low or below the certification level.

So far, US tests such as TEL and TUCE have been used for international comparison of economic literacy (see also Brückner and Zlatkin-Troitschanskaia, Chap. 6 in this volume). However, it is possible to overestimate or underestimate the levels of economic literacy by translation bias as shown by Hahn and Jang (2012). Such a finding invariably may harbingers more joint collaborations with interested countries to develop more uniform and standardized international tools.

References

- Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., Happ, R., Walstad, W. B., Yamaoka, M., & Asano, T. (2015). Gender effects in assessment of economic knowledge and understanding: Differences among undergraduate business and economic students in Germany, Japan, and the United States. *Peabody Journal of Education, 90*, 503–518.
- Buckles, S., & Walstad, W. B. (2008). The national assessment of educational progress in economics: Test framework, content specifications, and results. *Journal of Economic Education, 39*, 100–106.
- Council for Economic Education. (2010). *Voluntary national content standards in economics* (2nd ed.). New York: CEE.
- Hahn, K. (2013). Trends in economic education, 2005–2012: High schools and colleges. *Korean Journal of Economic Education, 20*, 157–174.
- Hahn, J., & Jang, K. (2010). Economic education in Korea: Current status and changes. *Journal of Economic Education, 41*, 436–447.
- Hahn, J., & Jang, K. (2012). The effects of a translation bias on the scores for the basic economics test. *Journal of Economic Education, 43*, 133–148.
- Jang, K., Hahn, K., & Kim, K. (2010). Comparative Korean results of TUCE with U.S. and Japan. In M. Watts, T. Asano, W. B. Walstad, & S. Abe (Eds.), *Comparative studies on economic education in Asia-Pacific region* (pp. 53–77). Yokohama: Shumpusa Publishing.
- Korean Educational Development Institute. (various years from 2011 to 2015). *Statistical year-book of education*.
- Maeil Business News Korea. (2013). *MK TEST questions previously appeared with explanations*. Seoul: MK Press.
- Ministry of Education. (2015). *National curriculum of social studies (in Korean)*. Sejong: MOE. Retrieved June 20, 2016, from http://www.moe.go.kr/web/100091/ko/board/view.do?bb_sId=141&pageSize=10¤tPage=0&encodeYn=N&boardSeq=60747&mode=view.
- Ministry of Education, Science and Technology. (2012). *National curriculum of social studies (in Korean)*. Seoul: MOEST.
- Moon, S., Kim, K., & Jang, K. (2010). A study on the Korea economic accreditation test. *Japanese Journal of Economic Education, 29*, 24–37.
- OECD. (2011). *Education at a glance 2011: OECD indicators*. OECD Publishing. <https://doi.org/10.1787/eag-2011-en>.
- Siegfried, J. J. (1979). Male-female differences in economic education: A survey. *Journal of Economic Education, 10*, 1–11.
- Song, S., Jang, K., & Hahn, K. (2015). Gender difference in economics knowledge understanding. *Korean Journal of Economic Education, 22*, 123–140.
- The MK TEST Web site: <http://exam.mk.co.kr>
- The TESAT Web site: <http://www.tesat.or.kr>
- Walstad, W. B. (2013). Economic understanding in U.S. high school courses. *American Economic Review, 103*, 659–663.
- Walstad, W. B., & Buckles, S. (2008). The national assessment of educational progress in economics: Findings for general economics. *American Economic Review, 98*, 541–546.
- Walstad, W. B., & Rebeck, K. (2008). The test of understanding of college economics. *American Economic Review, 98*, 547–551.
- Yamaoka, M. (2007). An international comparison of economics literacy of American and Japanese college students: A preliminary analysis of their understanding of college economics (7th Consumer Economics Test). *Bulletin of Graduate School of Asia Pacific Studies, 9*, 59–85.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education, 40*, 393–411.

Chapter 9

What Do We Know – What Should We Know? Measuring and Comparing Achievements of Learning in European Higher Education: Initiating the New CALOHEE Approach



Robert Wagenaar

Abstract Are the instruments to decide on the quality, suitability and relevance of higher education learning that we have at our disposal still adequate for today's dynamic world? Do students enrolled in higher education around Europe develop the competences they need? Are degree programmes delivering what they promise? Can we learn to compare student's achievements in different countries in a meaningful way? These very pertinent questions deserve an answer based on reliable evidence. This evidence is not at our disposal yet, although a set of tools has been developed in the framework of the Bologna Process that offers a good basis. However, the ultimate proof of the pudding is in the eating, which requires not only agreement on what should be but also on what has been learned. To respond to this need, the project *Measuring and Comparing Achievements of Learning Outcomes in Higher Education* (CALOHEE) has been established with support of the European Commission. CALOHEE is developing the instruments conditional for setting up transnational diagnostic assessments, which can be applied European-wide. CALOHEE delivers three types of outcome, outlined in this chapter: state-of-the-art reference points (benchmarks) for five academic sectors/subject areas, detailed assessment frameworks for these disciplines and a multidimensional assessment model that does justice to the mission and profile of individual higher education institutions and degree programmes.

The author thanks Ann Katharine Isaacs and Ingrid van der Meer for their critical remarks and suggestions for the improvement of a first draft of this chapter.

R. Wagenaar (✉)

University of Groningen, Groningen, The Netherlands

e-mail: r.wagenaar@rug.nl

9.1 Introduction

For the last 25 years, concern has been expressed about the suitability and relevance of higher education learning in today's dynamic world: first by the European Commission in several publications (EC 1991, 1997) and since 1998 by the European ministers of education, in the setting of the Sorbonne and the Bologna Declarations. Both have stipulated the need for reforms regarding the organization, design and implementation of degree programmes. Probably more than ever, the participation in and quality and performance of higher education are perceived as a significant factor for boosting economic growth and enhancing social well-being besides personal development (Katsarova 2015; see also Cain and Hearn in this volume). This not only applies to Europe (<https://www.whitehouse.gov/issues/education/higher-education/>). Living in a competitive world, there is a growing need felt to benchmark higher education performance at system level as well as at degree programme level (Katsarova 2015). Initiatives have been taken to set standards and develop indicators to allow for comparison (see also Coates in this volume). Until now, the most far-reaching one in terms of comparison at system level has been the Organisation for Economic Co-operation and Development (OECD) feasibility study *Assessment of Higher Education Learning Outcomes (AHELO)*. Although it resulted in three substantial volumes, the study did not produce a suitable and workable approach (Tremblay et al. 2012–2013). It showed however the challenges and limitations of global comparison of achievements of learning.

So far, more promising has been the work established in developing standards, descriptors and indicators. In 2005 the *European Standards and Guidelines for Quality Assurance* were published by the European Association for Quality Assurance in Higher Education (ENQA) which obtained wide approval and support. A decade later these were updated (ENQA et al. 2015). Another initiative in which much time and effort has been invested over the last 10 years or so is the development of 'overarching' or 'meta-level qualifications frameworks'. Good examples in this respect are the *Qualifications Framework for the European Higher Education Area* (QF for EHEA), based on the 'Dublin Descriptors', and the *European Qualifications Framework for Lifelong Learning* (EQF for LLL) (Bologna Working Group 2005; EC 2008). These have been complemented by national qualifications frameworks (CEDEFOP 2016). All these frameworks provide good indications of what is expected in terms of outcomes of a learning process at different levels. However, because of their purpose and role, the descriptors included in meta-frameworks are necessarily rather general.

Starting in 2001, benchmarks or reference points have been defined for specific subject areas or disciplinary fields, as well as for academic domains or sectors in the context of the Tuning Educational Structures in Europe projects and in European thematic networks (TNPs) (<http://www.unideusto.org/tuningeu/publications/subject-area-brochures.html>). These involved hundreds of academics from all over Europe. The European meta-frameworks and the Tuning subject area/sectoral qualifications frameworks should be perceived as complementary. Although they are

more detailed, subject area-based qualifications frameworks or benchmarks are also still rather general by nature, since each is expected to cover a broad academic field (see also Cain and Hearn in this volume). In their present forms, they are not suitable for precise measuring and comparing of learning. That requires more sophisticated frameworks. This chapter offers a new approach which is being developed in the framework of the feasibility study *Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe (CALOHEE)* (<https://www.calohee.eu>). In June 2016 a core objective of this study, the development of multi-dimensional ‘assessment frameworks’ at subject area level, was adopted as formal European Commission policy (EC 2016).

9.2 Developing an Infrastructure for Comparative Testing

With the Tuning projects aim to contribute to the realization of the main objectives of the Bologna Process, an important incentive for launching CALOHEE by the Tuning initiators has been the disappointing level of implementation of one of its current most important objectives for reform, the introduction of the concepts of active learning and the student-centred approach. Both were originally introduced in 2002 by Tuning. Recent research shows a disconnect between political ambitions and reality, that is, academic staff is unprepared and untrained for these concepts, and students are disappointingly unfamiliar with them (Gonzalez and Wagenaar 2003; Birtwistle et al. 2016). CALOHEE aims to serve as a source of inspiration for making the intended reforms a reality by offering a clear reference and the necessary approach and materials to facilitate the updating of content of degree programmes by making these ‘fitness in purpose’ and ‘fitness for purpose’. In other words, the outcomes of the learning process should meet the aims of the programme, as well as meet the needs and expectations of students and society, ensuring employment, personal development and civic, social and cultural engagement.

From this perspective the following questions were inspired: Do students enrolled in higher education around Europe develop the competences they need? Are degree programmes delivering what they promise? Can we learn to compare student’s achievements in different countries in a meaningful way? These are very pertinent questions given the amount of money involved in higher education for governments as well as for families and the students concerned. The notion of cost-benefit that is applied throughout society nowadays also applies to the higher education sector (see also Coates in this volume). In response to this, the ultimate aim of the CALOHEE initiative is to develop the infrastructure for setting up and implementing multidimensional assessments for five subject areas, chosen to represent five significant academic domains: humanities, social sciences, natural sciences, health care and engineering. These assessments and their underpinning frameworks should offer insight into whether the outcomes of learning match the investments made. The assessments for each of the five subject areas are intended to use a similar

methodology, but they shall be tailored to the characteristics of each field of studies, thus enabling a comparison of students' performance in a Europe-wide context.

The assessments will necessarily be multidimensional in order to allow for precise and fair measurement, taking into account the different missions, orientations and profiles of institutions and degree programmes. The outcomes of the assessments should not only offer institutions useful information to verify whether their students are achieving internationally defined standards of generic and subject-specific learning outcomes and are prepared sufficiently well for their role in society in terms of personal development, employability and civic, social and cultural engagement. Both the underpinning frameworks and the assessments intend to provide important information to the students themselves, so that they can understand better the objectives of their programmes and the competences they will gain and become proactive in the learning process. The frameworks and assessments will be designed in such a way as to stimulate academics to reform as well as to check that learning, teaching and assessment methods are truly aligned with the stated desired outcomes. Finally, the frameworks and the (outcomes of) assessments should play a key role in quality enhancement and assurance at degree programme level. Although actual comparative assessment is the ultimate aim, several steps have to be made first, which in itself is expected to offer a significant contribution towards the modernization and boosting of the quality and relevance of higher education programmes. These stepping stones involve the updating of existing subject area and sectoral qualifications frameworks as well as the development of meaningful 'assessment frameworks' which are drawn from these.

It is now widely accepted that both programme-level descriptors and unit- or module-level descriptors, described as programme and unit 'learning outcomes', are useful to determine whether the intended level of learning has actually been achieved. Experience has shown that learning outcome statements should be clearly and precisely formulated in order to guarantee objectivity/fairness and transparency. Tuning has developed a model, related to the work of educational scientists Bloom, Biggs and others (Lokhoff et al. 2010), which helps in elaborating reliable statements. Reliability is to be understood in this context as allowing for measuring and assessing the progress of learning and/or its achievement. The Tuning model distinguishes five elements that should be covered in a learning outcomes statement: verb, type, subject, standard and scope/context. Hence it is more precise than models which focus (mainly) on the use of the most appropriate 'verb' to indicate the level to be achieved during a specified piece of learning (Adelman 2015). Focusing on verbs has its limitations because it lacks precision in defining the scope and complexity and therefore the level of a learning outcome.

An additional instrument for determining the level of performance of an individual learner is so-called rubrics. Rubrics or score cards offer more detail and precision in terms of the criteria used to assess and grade a piece of student work and the weighting of different elements. Rubrics can have quite different formats and are used to assess an individual course unit or module. Although qualifications frameworks, level descriptors and rubrics are all indispensable tools for judging the quality of learning, they are not sufficient for comparing the results obtained by different

study programmes in the same field of study in a national or international context. This requires a completely new type of instrument being the mentioned ‘assessment framework’. Such a framework offers more detail than do qualifications frameworks about what a graduate in a particular subject area is expected to know, understand and be able to do when finishing his or her studies and/or a well-defined (structured) period of studies successfully. The European *Subject Area Assessment Framework* to be developed in the context of the CALOHEE feasibility study should thus provide a solid basis for constructing reliable and sustainable sets of assessment items for each of the five subject areas covered by the feasibility study.

9.3 Applied Principles

The OECD’s AHELO feasibility study has been inspirational for defining the CALOHEE study. On the basis of lessons learned, this has resulted in a completely different design. While AHELO was based on a top-down approach meant to find evidence regarding the performance of (national) systems, CALOHEE has chosen to use a bottom-up approach in order to give the academic community a central position in the further implementation of the process of modernization of higher education in Europe. It should offer ‘performance’ information at individual and aggregated at programme, institutional and national level. It involves 70 academics and 6 student representatives covering a wide range of countries. It builds on the work already carried out in the framework of the European Higher Education Area (Bologna Process) and the worldwide activities associated with Tuning. Although Tuning operates globally, all its projects are regionally based to do justice to cultural and other differences. In AHELO these differences were clearly underestimated despite a ‘contextual strand’ it had included in its outline, which should have offered a sufficient basis and safeguard to avoid the misinterpretation of results. In practice, AHELO struggled with insufficient cohesion.

What did not help either in this respect was the clear separation of a ‘generic skills and competences strand’ and two ‘subject specific knowledge and skills strands’ for, respectively, civil engineering and economics. In the philosophy of Tuning, these two strands cannot be separated and should be fully integrated in the teaching and learning process, based on the argument that generic competences are not only developed in the framework of a domain of knowledge but are also perceived differently between educational sectors. Another serious weakness proved to be that the design did not allow for differences in missions and profiles of higher education institutions, for example, more research driven and more applied formats. The response of CALOHEE to this diversity is the application of a multidimensional approach by using two main parameters for assessment: ‘knowledge: theoretical and methodology’ and the ‘application of knowledge and skills’. Taking into account the responsibility of higher education to prepare graduates for their role in society, another two categories or parameters have been added to the two mentioned which were not covered by AHELO: preparation for employability and civic, social and cultural

engagement. This makes CALOHEE much more comprehensive and relevant as reference for what ‘should’ be learned according to different stakeholder groups.

A further innovation introduced by CALOHEE is the merging of the two existing European meta-frameworks in one model and the introduction of the concept of dimensions, covering areas of learning (Wagenaar 2013). This model is also applied to organize the descriptors of competences in one encompassing table or grid per level (Ba and Ma/EQF 6 and 7) which should represent a so-called meta-profile for the sector and the subject area involved. These tables are a crucial addition to the existing Tuning subject area reference points brochures which have been published since 2008. Another new element to these brochures is the identification of roles and tasks of graduates which go beyond an inventory of occupations. To collect more detailed information, a questionnaire was distributed among the academics involved in CALOHEE, the outcomes of which have proven to be an eye-opener. The outcomes showed that it is indeed possible to identify clear accumulated sets of tasks and roles of graduates per subject area which offer much more useful information to take into account when designing and updating the content of degree programmes than an overview of typical occupations can offer. From the material available, it is obvious that many of these typical roles and tasks are not ‘trained’ (very well or explicitly) during higher education programmes (<https://www.calohee.eu/>).

Having the principles outlined above, the remainder of this chapter will provide insight into (1) the definition of the assessment framework proposed; (2) the application of qualifications frameworks and so-called dimensions to construct an assessment framework; (3) the multidimensional parameters identified, that is, the items to be assessed, in terms of theory, methodology, skills, application, employability and civic-related competences; and (4) the structure of the framework, that is, the topics of assessment and their related possible learning, teaching and assessment approaches.

9.4 Assessment Framework Definition

The term ‘assessment framework’ can have different meanings. On the one hand, it may refer to an instrument used as a basis for an accreditation procedure, that is, to check whether a study programme meets minimum quality standards (ECA 2014). On the other, it can also be understood as a framework, which offers a detailed scheme or schedule of phases in an assessment process, including the different approaches to be used with respect to the course units/modules that form a particular study programme (<https://www.nottingham.ac.uk/teaching/assessmentfeedback/assessmentframework.aspx>). The teaching staff involved in such a programme is expected to respect this scheme when implementing the programme. It should offer a well-thought, thorough and balanced structure for assessment of the different programme components.

In the case of CALOHEE, ‘assessment framework’ has a third meaning. It is a table which contains the learning outcomes or descriptors defined as part of a sub-

ject area qualifications framework and more precise subsets of each one of them. Each subset, taken together, describes in some detail the key elements and topics covered by a learning outcome statement. In addition, the assessment framework intends to offer insight in the most appropriate strategies and approaches to assessing the constituent elements of each learning outcome. The term is used in CALOHEE in the same way as in the OECD AHELO feasibility study, where assessment frameworks were defined for the disciplinary fields of Economics and Civil Engineering, based on the respective Tuning AHELO conceptual frameworks for those two subject areas (OECD 2011a, b, 2012a, b).

9.5 Qualifications Frameworks and Dimensions

As mentioned above, the assessment frameworks to be developed will be based on the grids or tables of descriptors included in the Tuning sectoral and subject area qualifications frameworks. The EQF for LLL uses the categories of knowledge, skills and competences to structure its descriptors. Thus, the three columns form in CALOHEE terms a ‘knowledge framework’, a ‘skills framework’ and a ‘competency framework’, linked by level. The last column, the ‘competency framework’, refers to the world of work and society and identifies the competences required to operate successfully in the work place and as an active citizen. In the EQF, the competency column builds on the other two elements: knowledge and understanding and the skills necessary to develop and use this knowledge. Together these can be seen as ‘content-related competences’ or ‘subject-specific competences’. As is well known, besides these, Tuning distinguishes ‘generic or general competences’, which are grouped in three categories: instrumental, interpersonal and systematic competences. These should be covered in the ‘competency’ strand but are also related to the ‘skills’ strand.

To illustrate this point, it is worth mentioning that over time many competency frameworks have been developed for a specific job sector, company or institution. These define the requirements for a given job and are used in job vacancy announcements. These announcements normally contain content-related or subject-specific competences as well as generic competences. As an example of a well-developed competency framework, we may take the one the OECD produced in 2014 for the selection/assessment and promotion of its own staff (OECD 2014). This competency framework is linked to the catchwords: learn, perform and succeed. It makes a distinction between ‘technical competences’ (subject-specific competences) and ‘core competences’ (generic competences). It identifies 15 ‘core competences’ which are organized in three clusters: ‘delivery-related competences’ focusing on achieving results, ‘interpersonal competences’ focusing on building relationships and ‘strategic competences’ focusing on planning for the future. The ‘delivery-related competences’ are analytical thinking, focus on achievement, drafting skills, flexible thinking, resource management, teamwork and team leadership. The interpersonal competences selected are client focus, diplomatic sensitivity, negotiation

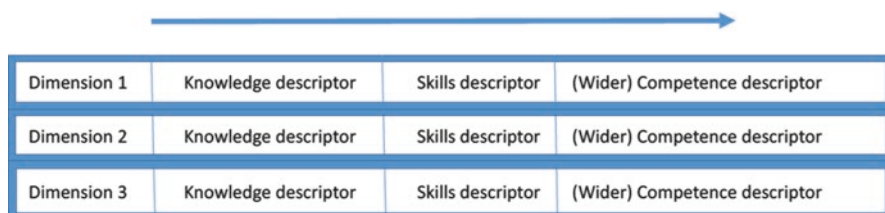
and organizational knowledge. The strategic competences identified are developing talent, organizational alignment, strategic networking and strategic thinking. For each of these competences, a definition was formulated.

Based on these competences, the OECD competency framework offers indicators for different levels, which are associated with types of jobs. Level 1 is typically associated with jobs as assistants, secretaries and operators and the like; level 2 with jobs as statisticians, corporate management and administration assistants/officers, logistics officers and documentalists; level 3 with jobs as economists/policy analysts, IT analysts and human resources advisers; and level 4 with jobs as senior economists/policy analysts or managers. Level 5, the highest level identified, is associated with jobs as heads of division, counsellors, deputy directors and directors and so forth. The typical jobs identified for the OECD might have limited value for many of the subject area covered by CALOHEE, but the operationalization of levels is useful. This is because the indicators used are clearly related to levels of responsibility and autonomy, the main indicators covered in the 'competence strand' of the EQF. The OECD framework is also relevant because it makes a clear link to the 'tasks and roles' executed as part of the jobs identified. The OECD document distinguishes three job families: 'executive leadership', 'policy research, analysis and advice' and 'corporate management and administration'. The OECD framework is only one example; many others can be found on the Internet (<https://www.microsoft.com/en-us/education/training-and-events/education-competencies/default.aspx?tabselect=1>). Besides in the management sector, competency frameworks have been drawn up and are applied in the health-care sector (Sastre-Fullana et al. 2014). Besides these job-related frameworks, recently a competency framework has been published for student work-based learning covering all levels of higher education, including the PhD (Jones and Warnock 2014).

As stated above, for the purposes of the CALOHEE study, the EQF for LLL has been merged with the QF for EHEA to make use of 'the best of two worlds'. While the EQF is focused on the application of knowledge and skills in society, the focus of the QF for the EHEA is more related to the learning process itself: it applies descriptors which cover different areas or 'dimensions' of learning: knowledge and understanding, application of knowledge and understanding in relation to problem solving, making judgments, communicating information, conclusions, etc. and learning capability. In developing the CALOHEE approach, the conclusion has been drawn that 'dimensions' are indispensable to define the field of study for which it is required to distinguish the different constituting areas. The 'dimension approach' is complementary to the three categories included in the EQF for LLL. Dimensions help give structure to a particular sector or subject area and also make these more transparent. The use of 'dimensions' facilitates breaking down the rather general level descriptors into more precise ones. This process is necessary in order to develop an assessment framework, which must be sufficiently detailed to permit comparing and measuring. Such an approach also provides far better indicators for evaluating the quality of a degree programme than are available at present. Initially, a number of the academics involved in CALOHEE expressed their doubts

about the usefulness of the proposed merging of the two frameworks, but at its second general meeting taking place mid-November 2016, it was unanimously concluded this is the best way forward.

Although there should be an obvious connection with the five or six areas of learning (depending on the cycle covered) or dimensions formulated as general descriptors in the QF for the EHEA, each sector must define its own set of sectoral/subject area dimensions in order to be able to do justice to its field. In the sectoral frameworks developed so far, diversity has been found between sectors as well as some overlap. Each dimension in a Tuning CALOHEE Qualifications Framework includes three *related* descriptors, respectively, for knowledge, skills and (wider) competences. This is illustrated in the Fig. 9.1.



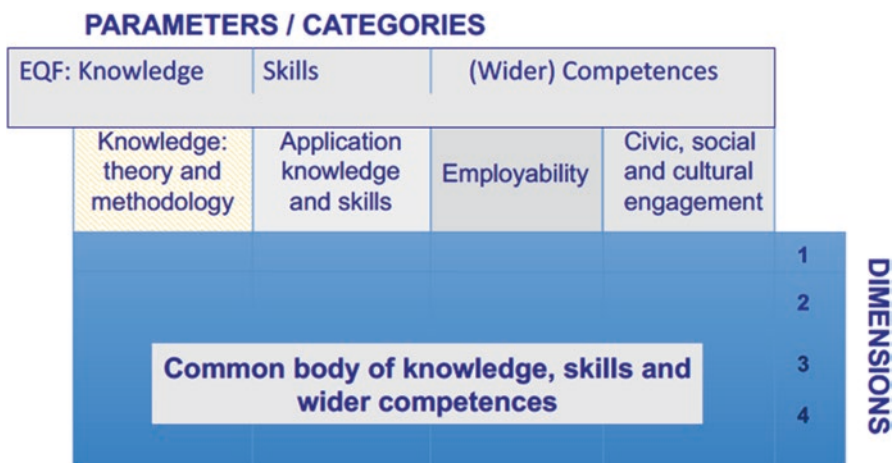
Dimension 1	Knowledge descriptor	Skills descriptor	(Wider) Competence descriptor
Dimension 2	Knowledge descriptor	Skills descriptor	(Wider) Competence descriptor
Dimension 3	Knowledge descriptor	Skills descriptor	(Wider) Competence descriptor

Fig. 9.1 Dimensions in a tuning CALOHEE qualifications framework

The ‘skills descriptor’ builds on the ‘knowledge descriptor’ and the ‘(wider) competence descriptor’ on the other two. In Tuning and CALOHEE the term ‘wider competences’ is preferred, because it takes into account the fact that knowledge and understanding must also be understood as competences, in this case ‘subject-specific’ ones or in OECD terms ‘technical competences’. Using the term ‘wider competences’ also expresses the fact that the aim of a period of study is both to foster personal development and to increase the learner’s competences for future employment.

9.6 Multidimensional Parameters

In order to accommodate the different missions and profiles of higher education institutions and their programmes, the CALOHEE assessment frameworks will be based on four parameters or categories. This is completely compatible with the existing Tuning CALOHEE sectoral/subject area qualifications frameworks whose core is formed by the grid or table of descriptors/learning outcomes. As the Fig. 9.2 illustrates, the four parameters of assessment are related to the three strands: ‘knowledge’, ‘skills’ and ‘(wider) competences’. The last strand is split into two: employability and civic, social and cultural engagement. The term ‘active citizenship’ is avoided, because it has a negative connotation in large parts of Europe.



Assessment framework

Fig. 9.2 Multidimensional parameters of the assessment framework

The distinction in strands is made for reasons of clarity, although it must be kept in mind that the four strands are closely interrelated, as are the three strands in the EQF for LLL and the five or six dimensions in the QF for the EHEA.

The first parameter encompasses the core knowledge of a particular academic field as well as the related theoretical concepts and methodologies which are judged essential for a good understanding of that field. The depth to which this knowledge and its understanding are developed in a programme depends on the type of degree programme and type of institution offering it. For example, in the case of a research-intensive institution, deep knowledge of theoretical concepts and methodologies in relation to highly developed analytical competences/skills and critical thinking will be considered essential. While the outcomes of the Tuning surveys have shown that stakeholders consider the ability to apply knowledge and skills in practice – the second strand – very important in preparing for a societal role, in the case of the research-intensive institution the focus will be much stronger on the first strand. The balance will be different in the case of a university of applied science or a more applied degree programme. However, the CALOHEE assessment framework will indicate the optimum achievement level in both categories (for both BA and MA), that is, the highest level achievable and feasible for a higher education degree programme.

This means that students are not all expected to achieve the highest levels which are formulated as ‘intended’ learning outcomes in the framework. The norm of achievement – threshold, average, above average, excellent – with regard to each of the parameters will depend on the type of programme taken by the student, as well as its aims. This approach, which can be compared to the tests used to select pupils/students for different types of secondary and higher education, does justice to CALOHEE’s multidimensional approach. It also takes into account that in national and international contexts, a distinction is made between more and less ‘prestigious’

universities or other types of higher education institutions if these exist (grand écoles, skola normal, etc.). Although all these institutions will offer bachelor and/or master programmes (or their equivalents), it does not mean that these are understood to be of the same higher education ‘type’ or ‘level’. This is why it so important to distinguish profiles and missions of institutions, each of which have an intrinsic value and place and role in the higher education landscape but therefore also have the obligation to describe and justify the choices they make.

Once the ‘optimum’ feasible learning outcomes are defined, it is essential to make subdivisions which reflect the different profiles of higher education institutions and programmes in an appropriate manner. These should also be the basis for deciding the norms to use when comparative assessments are organized. In order to avoid complicating the model excessively, it is proposed to develop two main subdivisions (research based and applied), which can be further split into two subsets, so as to distinguish level. This would provide grids for four types of degree programmes, having partially different programme learning outcomes and taking into account more academic and more professional orientations. All types, however, are expected to cover the identified common body of knowledge, skills and (wider) competences, and all students are expected to meet a threshold level to be identified and agreed upon by the academic communities responsible.

The parameter related to employability has already been discussed above by linking it to competency frameworks. As the OECD example shows us, different programme profiles might lead to different types of jobs given the tasks and roles related to these jobs which require different levels of competence. Employability can be defined in short as the skills and abilities that allows someone to be employed. The United Kingdom (UK) Higher Education Academy/Enhancing Student Employability Coordination Team (ESECT) have come up with the following definition of employability-related competences: *A set of skills, knowledge and personal attributes that make an individual more likely to secure and be successful in their chosen occupation(s) to the benefit of themselves, the workforce, the community and the economy* (York 2006). It is obvious that both subject-specific and general/generic competences are understood to be quite important in this context. In this last respect, the publication of the UK Higher Education Academy *Student employability profiles* is of relevance. It offers short profiles for each of the subject areas covered in the CALOHEE project (Rees et al. 2006).

Given the role of higher education institutions to prepare students for their role in society and to form strong bases for personal development, in addition to preparing them for participating in the work force, CALOHEE holds that it is important – even essential – that attention in the learning process is paid to civic, social and cultural engagement. This formulation is often referred to in the European context as ‘active citizenship’. It may well be that this aspect is not explicitly pursued at present in the vast majority of higher education programmes, but this is a serious omission, given the fact that the stability of many societies is under severe pressure. Interrelated challenges such as the refugee crises, the lasting effects of the 2008 financial crisis, the rapidly changing geopolitical context, the negative consequences of globalization, xenophobia, populism and most recently the Brexit and United

States (US) Presidential election, which reflect all these elements, shake the foundations of societies and their constituent components.

It is expected that the competences reflected in this strand will be largely the same for all subject areas, although the perception of their importance can differ. For academic fields such as history, educational sciences and teacher training, their 'weight' in the curriculum might be greater than in other disciplines. Recent publications show there is global attention for this category of learning. In 2010 the Australian government published its *Civics & Citizenship Education Professional Learning Package* (Australian Government 2010), and although it was meant for secondary education in particular, the topics covered seem to be relevant for higher education as well. It offers three modules to foster 'civics and citizenship', respectively, 'in the classroom', 'beyond the classroom' and 'participation in the community'. The focus in the modules is on 'civics and citizenship education knowledge, skills and dispositions' (an artificial habit, a preparation, a state of readiness or a tendency to act in a specified way that may be learned).

Probably even more important in the CALOHEE context is the 2016 publication of the Council of Europe, *Competences for Democratic Culture: Living together as equals in culturally diverse democratic societies* (Council of Europe 2016). In the publication 20 generic competences are distinguished, which are clustered in four groups: values, attitudes, skills and knowledge and critical understanding. By values is meant human dignity and human rights, cultural diversity, valuing democracy, justice, fairness, equality and the rule of law. The label attitudes encompass openness to cultural otherness and to other beliefs, world views and practices as well as civic-mindedness, responsibility, self-efficacy and tolerance of ambiguity. As skills have been identified autonomous learning, analytical and critical thinking, listening and observing, empathy, flexibility and adaptability, co-operation, conflict-resolution and linguistic, communicative and plurilingual abilities. The knowledge category lists knowledge and critical understanding of the self, knowledge and understanding of language and communication as well as the world, in terms of politics, law, human rights, culture, cultures, religions, history, media, economies, environment and sustainability.

From this list it is obvious that competences relevant for employability overlap with those for civic engagement. It shows that combining both employability and civic, social and cultural engagement in the 'wider competences' parameter/category is a sensible solution. The list of 20 generic competences chosen by the Council of Europe is based on a longer list of 55 identified in 101 competences schemes. Each of the 20 competences is clarified in the document and supported by a number of pre-assumptions, ranging from 3 to 12 statements. They offer clarity about what is expected of a citizen in a democratic culture. Taken together, these statements should be measurable.

A Educational Testing Service (ETS) research group also has studied the issue. The report by Judith Torney Puta et al. (2015) *Assessing civic competency and engagement. Research Background, Frameworks, and Directions for Next-Generation Assessment* stresses that civic learning is increasingly recognized as being important by both the higher education sector and workforce communities. It offers a review of the outcomes of some 30 projects covering 'existing frame-

works, definitions and assessments of civic related-constructs'. It identifies 31 competences ranging from civic literacy, civic engagement, civic identity, political knowledge, civic knowledge and skills, ethical and social responsibility in a diverse world, civic-mindedness and civic responsibility to political and civic participation. It also addresses the term 'civic learning' in terms of learning outcomes in the Lumina US Degree Qualifications Profile (DQP) both at associate level (level 5 of the EQF) and at bachelor level (<http://www.luminafoundation.org/files/resources/dqp.pdf>). The study offers a table of 'existing assessments measuring civic competency and engagement' and comes up with its own framework, distinguishing between the civic competency domain (covering civic knowledge, analytical skills, participatory and involvement skills) and the civic engagement domain (covering motivations, attitudes and efficacy, democratic norms and values and participation and activities). These competences are defined and completed with measurable topics/learning outcomes. The report concludes with examples of so-called 'test item formats' to assess civic competency and engagement.

These publications – together with others (<http://compact.org/resource-posts/assessment-of-students-civic-learning-and-development/>, Council of Europe 2016) – have offered a good basis to give substance to the parameter of assessment and allowed for defining concrete learning outcomes, which can be learned, taught and measured. It has resulted in a CALOHEE model – based on an analysis of present developments and the recent literature mentioned above – that contains four dimensions:

- Societies and cultures: Interculturalism and conflict management
- Processes of information and communication
- Processes of governance and decision-making
- Ethics, norms, values and professional standards

For each of these dimensions, knowledge, skills and (wider) competence descriptors have been defined. It is expected that these are integrated in the qualifications frameworks of each subject area (CALOHEE 2017).

9.7 Topics of Assessment, Teaching and Learning

Keeping the proposed four parameters, strands, dimensions and the main subdivision and its subsets in mind, the first step is to break down each of the descriptors linked to the 'dimensions'-related knowledge, skills and (wider) competences. Only after their breakdown has been realized does it seem feasible to give substance to the subdivision subsets as identified.

The splitting-up can be accomplished by identifying the different components which make up these descriptors. It has been suggested to distinguish 3–5 components to be formulated as subsets/sub-descriptors of each dimension. The lists of 'subject-specific competences' and 'general or generic competences' which have been identified by each Tuning subject area group as being the most relevant for the academic field (sector and subject area) should serve as a basis. The breakdown can be visualized as follows (Fig. 9.3).

Dimension	Knowledge descriptor	Skills descriptor	(Wider) Competence descriptor
1.	Sub-descriptor 1-1	Sub-descriptor 1-2	Sub-descriptor 1-3
2.	Sub-descriptor 2-1	Sub-descriptor 2-2	Sub-descriptor 2-3
3.	Sub-descriptor 3-1	Sub-descriptor 3-2	Sub-descriptor 3-3
4.	Sub-descriptor 4-1	Sub-descriptor 4-2	Sub-descriptor 4-3
5.	Sub-descriptor 5-1	Sub-descriptor 5-2	Sub-descriptor 5-3




Fig. 9.3 Descriptors linked to the ‘dimensions’-related knowledge, skills and (wider) competences

To make this model more concrete, the following (provisional) example at level 6 (bachelor) is taken from the work done by the CALOHEE Subject Area Group of History. The following dimensions for humanities/history are distinguished: ‘the human being: cultures and societies’, ‘texts and contexts’, ‘theories and concepts’, ‘interdisciplinarity’, ‘communication’, ‘initiative and creativity’ and ‘professional development’. This results in the following Table 9.1.

Table 9.1 Overview of the dimensions

Dimension	Knowledge	Skills	Competences
Human beings: cultures and societies L6_1. Level descriptor	Demonstrate basic knowledge and critical insight into changes and continuities in the human condition, environment and experience, in institutions and modes of expression, ideas and values in a diachronic perspective	Drawing on knowledge of history, identify and define, with guidance, significant problems and areas of enquiry with respect to social and cultural interaction	Apply historical knowledge and perspectives in addressing present-day issues, bringing to bear analytical understanding and respect for the individual human being in his/her personal, cultural and social dimension
Texts and contexts L6_2. Level descriptor	Demonstrate knowledge and understanding of the main kinds of sources for historical research	Identify, select with guidance and present information from a variety of historical sources in an appropriate form	Retrieve, manage and use information in order to formulate and address problems in their contexts using suitable methodologies

(continued)

Table 9.1 (continued)

Dimension	Knowledge	Skills	Competences
Theories and concepts L6_3 Level descriptor	Collect knowledge about and classify a range of analytical, theoretical and methodological approaches relevant to history. Demonstrate orientation in the major themes of present historical debate and knowledge of world chronology	Apply appropriate critical and methodological approaches to historical questions	Examine and explore societal issues and processes using relevant theories and concepts
Interdisciplinarity L6_4 Level descriptor	Demonstrate knowledge of the intellectual underpinnings and contexts of history in relation to other fields of study	Utilize, when opportune, knowledge and understanding from other fields to address problems and issues in the historical domain	Work with others in a multidisciplinary and/or multi-national setting when useful
Communication L6_5 Level descriptor	Demonstrate knowledge of the main means of communication used to convey information and perspectives in both academic and broader public contexts	Write and speak correctly in one's own language according to the various communication registers (informal, formal, scientific). Understand the appropriate terminology and modes of expression of the field of history also in a second language	Demonstrate ability to listen, understand different viewpoints and discuss ideas, problems and solutions with diverse audiences
Initiative and creativity L6_6 Level descriptor	Demonstrate knowledge of the ongoing nature of historical research and debate and of how historians contribute to key areas of academic and public discussion	Approach issues with curiosity, creativity and critical awareness; retrieve and handle information from a variety of sources (electronic, written, archival, oral) as appropriate to the problem, integrating it critically into a grounded narrative	Reflect on one's own perspective, capabilities and performance to improve and use them in a creative way. Think in scientific terms, pose problems, gather and analyse data and propose findings
Professional development L6_7 Level descriptor	Demonstrate knowledge of the intellectual bases and ethical aspects of historical studies and of the diverse contributions historians make to society	Methods to stay up to date with learning. Work autonomously and in a team, taking initiatives and managing time	Identify and/or create an appropriate study and/or work environment and participate effectively in it

Table prepared by the members of the CALOHEE Subject Area Group of History, July 2017

The first dimension, human beings: cultures and societies, which is meant to act as an overarching one for the field of history and the sector of Humanities is used here as an illustration of the outcomes a breakdown of a dimension in sub-dimensions or subsets offers (Table 9.2).

Table 9.2 Dimension 1: Human beings: cultures and societies

EQF level 6 (bachelor)	Knowledge	Skills	Competences
L6_1. Level descriptor	Demonstrate basic knowledge and critical insight into changes and continuities in the human condition, environment and experience, in institutions and modes of expression, ideas and values in a diachronic perspective	Drawing on knowledge of history, identify and define, with guidance, significant problems and areas of enquiry with respect to social and cultural interaction	Apply historical knowledge and perspectives in addressing present-day issues, bringing to bear analytical understanding and respect for the individual human being in his/her personal, cultural and social dimension
Subset 1	Show general acquaintance with diverse criteria of historical explanation and understanding on different time and spatial scales. Demonstrate awareness of how explanations and interpretations are conceptualized	Formulate historical explanations and interpretations of phenomena and processes through comparison and differentiation using quantitative and qualitative methods	Recognize consistent interrelations concerning phenomena and processes of different nature and scale, at the same time showing awareness of their uniqueness
L6_1.1 Historical interpretation of changes and continuities			
Subset 2	Relate social and economic change to environmental transformations and to the accumulation/ modification of knowledge	Describe the interaction between the natural environment and social change, on the one hand, and knowledge production on the other	Evaluate the impact of knowledge production and accumulation on society and the environment and vice versa
L6_1.2 Environmental transformations and knowledge development			
Subset 3	Demonstrate knowledge about power relations and how they shape collective organizations, institutions and representations of the world through conflict, negotiation and adaptation	Recognize tools and mechanisms of power in societal and collective relations and their genesis, continuity and transformations in time	Contribute to discussions and debates on power relations and political organization in a broad sense, placing them in historical perspective
L6_1.3 Power relations and organization			

(continued)

Table 9.2 (continued)

EQF level 6 (bachelor)	Knowledge	Skills	Competences
Subset 4 L6_1.4 Religious beliefs and practices	Demonstrate knowledge about modes of expression and transmission of beliefs and practices concerning moral values, immaterial and transcendental concerns and narratives and their dynamics	Describe different conceptual frameworks, symbolic representations and discourses that underpin and support collectively held beliefs and related practices	Engage critically with the dynamics of collective beliefs and practices and how they are expressed by individuals and groups
Subset 5 L6_1.5 Intercultural encounters	Demonstrate knowledge about intercultural encounters and their consequences on every field of human activities and on personal and collective identities	Describe and illustrate different dimensions (e.g. social, economic, religious and political) in cultural encounters via comparison and connections of specific cases	Contribute to understanding and respect for individuals and groups in their personal, cultural, economic and political and social dimension; conduct critical appraisal of conflicting views and facilitate intercultural mediation

Table prepared by the members of the CALOHEE Subject Area Group of History, July 2017

Each sub-descriptor describes – in the form of a learning outcomes statement – a core element or topic constituting the respective ‘knowledge descriptor’, the ‘skills descriptor’ and the ‘wider competence descriptor’. These sub-descriptors can be compared to the learning outcomes statements as defined for the ‘highest’ of a range of successive units or modules in a degree programme (a so-called ‘learning string’), defining the level to be achieved. The sub-descriptors have to be formulated in such a way that they can not only be measured but also be learned and taught. Like descriptors, sub-descriptors should be appropriate for the cycle (BA and MA) for which they are defined. However, as in the case of the cycle-level descriptors, it is advisable (if feasible and suitable) to develop these at the same time, to secure a fair balance. When formulating the sub-descriptors, it is suggested to keep the Tuning model for defining learning outcomes in mind (Lokhoff et al. 2010; Moon 2002; EC 2015).

As part of the process of defining a sub-descriptor, it is thought necessary to identify the appropriate learning, teaching and assessment approaches, methodologies and techniques. This can be done at the level of the descriptor as long as all sub-descriptors can be covered. Experience of linking specific approaches to learning, teaching and assessment to descriptors has already been successfully applied in a recent Tuning project, TuCAHEA, focussing on Central Asian countries, although not in as much detail as is proposed here (<http://www.tucahea.org>). To obtain a more up-to-date overview of the current approaches applied, questionnaires were distributed regarding modes of teaching and learning and on modes of assessment among the CALOHEE membership. It was also asked to identify modes of assessment to

‘measure’ competence development for a set of key generic competences. The outcomes show that still mostly rather traditional assessment forms are applied. As far as the assessment of generic competences is concerned, the questionnaire shows a rather ambiguous picture because the respondents have no clear ideas on which modes could be best applied. This confirms earlier findings that the student-centred approach has not been implemented widely yet. It is relevant to mention here that of the 101 respondents, 97% confirmed that their institution is representative for their country, as is 93% of their degree programmes (CALOHEE 2016a, b). It shows the need for examples of ‘good practice’ to be identified by the subject area groups as part of the process of updating the present Tuning reference points brochures. The interrelation between descriptors, sub-descriptors and approaches for learning, teaching and assessment is shown below (Fig. 9.4).

Dimension	Knowledge descriptor	Skills descriptor	(Wider) Competence descriptor
1.	Sub-descriptor 1-1	Sub-descriptor 1-2	Sub-descriptor 1-3
1a	Learning approach	Learning approach	Learning approach
1b	Teaching approach	Teaching approach	Teaching approach
1c	Assessment approach	Assessment approach	Assessment approach
2.	Sub-descriptor 2-1	Sub-descriptor 2-2	Sub-descriptor 3-3




Fig. 9.4 Interrelation between descriptors, sub-descriptors and approaches for learning, teaching and assessment

Not every key element or topic described in a sub-descriptor has to be covered by each degree programme. Whether and to which level each will be covered in practice will depend on the profile and mission of the programme concerned.

9.8 Outcome of the Exercise

The outcome of the exercise will be an assessment framework for the subject area covering both first and second cycle (bachelor and master). Based on the dimensions identified, it will contain ‘knowledge descriptors’, ‘skills descriptors’ and ‘wider competences descriptors’, all of which will be underpinned by more precise sub-descriptors. Each sub-descriptor formulated as a learning outcome will cover a core element or topic. For each sub-descriptor or combination of sub-descriptors learning, teaching and assessment approaches will be identified. These should allow for the achievement of the learning outcome(s) and be presented as examples of good practice. It is not considered sufficient in this respect just to mention a method or approach, rather it is necessary to indicate ‘why’ this approach or method is used and ‘how’ it is applied in addition to the ‘what’ described in the learning outcome.

An assessment framework containing these elements will not only serve as an important reference for constructing new programmes and modernizing, revising and enhancing existing ones but will also serve as a fair indicator for the completeness and (high) quality of a degree programme allowing for different missions and profiles. But most of all, it will be a reliable instrument for measuring and comparing the achievement of learning outcomes in a national and international setting and therefore will act as a sustainable basis for making a next step: the development of the actual measurement instrument, that is, sets of consistent test formats and items.

In AHELO two more traditional formats were applied for assessment: multiple choice tests and constructive response tests of which the last required manpower-based assessments. For reasons of reliability, efficiency and cost-effectiveness, CALOHEE strives for machine-based testing only. Conditional is that this type allows for the assessment of profound knowledge and understanding as well as high-level skills. One should think of critical awareness, analysing and composition skills, for example. This implies that formats should be developed and applied which make it possible to facilitate text interpretation and analysis but also to identify best strategies and methodologies for solving a problem. This will require the application of new forms of (statistical) measurement methods and validation approaches for assessments, which are still in the process of development (e.g. Shavelson et al. in this volume). It is expected that the use of algorithms will revolutionize computerized assessments. It can build on forms already available, such as responding to and analysing footage and computer simulation. Also, strategic computer games technology can be of service here. Given the speed at which technology is developing, the perspectives are quite promising and will allow for forms of comparative measuring not many could foresee almost a decade ago when AHELO was launched. That this is possible no longer seems to be an issue, but rather the question that remains is *when* it will become possible to find confirmation in what we really should know as the outcome of a process of learning, being much more relevant than what we already do know.

References

- Adelman, C. (2015, February). *To imagine a verb: The language and syntax of learning outcome statements* (Occasional Paper No. 24). Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. <http://www.learningoutcomeassessment.org/documents/OccasionalPaper24.pdf>.
- Australian Government. (2010). *Civics & citizenship education professional learning package*. http://www.civicsandcitizenship.edu.au/verve/_resources/DEEWR_CCE_PLP.pdf
- Bologna Working Group on Qualifications Frameworks. (2005, February). *A framework for qualifications of the european higher education area*. Copenhagen: Danish Ministry of Science, Technology and Innovation.
- Birtwistle, T., Brown, C., & Wagenaar, R. (2016, May). A long way to go ... a study on the implementation of the learning-outcomes based approach in the EU. *Tuning Journal for Higher Education*, 3(2), 429–462.
- CALOHEE. Website. <https://www.calohee.eu>

- CALOHEE. Summary of CALOHEE questionnaires outcomes on typical degrees and occupations (2016a). Retrieved November 29, 2016, from <https://www.calohee.eu/summery-calohee-questionnaires-typical-degrees-occupations/>
- CALOHEE. (2016b). *Outcomes questionnaire modes of assessment*. Meeting document Second General Meeting, Porto, 18–19 November 2016, 9–40. Groningen.
- CALOHEE. (2017). CALOHEE working paper for *Civic, social and cultural engagement*, Groningen.
- Campus Compact, Assessment of Students' Civic Learning and Development. <http://compact.org/resource-posts/assessment-of-students-civic-learning-and-development/>
- CEDEFOP. (2016). *Overview of national qualifications frameworks. Development in Europe* (Thessaloniki). Retrieved November 29, 2016, from <http://www.cedefop.europa.eu/en/publications-and-resources/publications/8606>
- Council of Europe. (2016). *Competences for democratic culture: Living together as equals in culturally diverse democratic societies* (Strasbourg). http://www.coe.int/t/dg4/education/Source/competences/CDC_en.pdf
- ECA. (2014). *Assessment frameworks for joint programmes*. Retrieved October 19, 2016, from http://ecahe.eu/w/images/e/e6/Assessment_framework_for_joint_programmes_in_single_accreditation_procedures_-_ECA.pdf
- ENQA, ESU, EUA, EURAHE. (2015). *Standards and guidelines for quality assurance in the European Higher Education Area (ESG)*, Brussels.
- European Commission. (1991). *Memorandum on higher education in the European community*, Luxembourg.
- European Commission. (1997). *Communication from the European commission: Towards a Europe of knowledge*, Luxembourg.
- European Commission. (2008). *The European qualifications framework for life long learning (EQF)*, Luxembourg.
- European Commission. (2015). *ECTS users' guide 2015*, Brussels.
- European Commission. (2016). *Communication from the commission to the European parliament, the council, the European Economic and Social Committee and the Committee of the Regions. A new skills agenda for Europe. Working together to strengthen human capital, employability and competitiveness*, (COM 381/2) (Strasbourg).
- Gonzalez, J., & Wagenaar, R. (Eds.). (2003). *Tuning educational structures in Europe. Final Report Phase One*, Bilbao and Groningen.
- Greater London Authority, Competency Framework. Guide for managers and staff. Retrieved November 28, 2016., from https://www.london.gov.uk/sites/default/files/competency_framework_guidelines_0.pdf.
- Jones, H. M., & Warnock, L. (2014, September). *Towards a competency framework for student work-based learning*. York: The Higher Education Academy. Retrieved November 28, 2016, from <https://www.york.ac.uk/media/biology/Towards%20a%20competency%20framework%20for%20student%20work-based%20learning.pdf>
- Katsarova, I. (2015, March). *Higher education in the EU. Approaches, issues and trends. In-depth analysis* (European Parliamentary Research Service, March 2015), 1 and 23–28.
- Lokhoff, J., Wegewijs, B., Durkin, K., Wagenaar, R., Gonzalez, J., Isaacs, A. K., Dona della Rose, L., & Gobbi, M. (2010). *A tuning guide to formulating degree programme profiles. Including programme competences and programme learning outcomes*. Bilbao/Groningen/The Hague.
- Lumina Foundation. (2014, October). *Degree Qualifications Profile (DQP). A learning-centered framework for what college graduates should know and be able to do to earn the associate, bachelor's or master's degree*, Indianapolis.
- Microsoft, Microsoft Education competencies for teachers and school leaders. Retrieved October 18, 2016., from <https://www.microsoft.com/en-us/education/training-and-events/education-competencies/default.aspx?tabselect=1>
- Moon, J. (2002). *The module and programme development handbook*. London: Kogan Page.
- OECD. (2011a). *Tuning-AHELO conceptual framework of expected and desired learning outcomes in economics* (OECD Education Working Papers, No.59) Paris: OECD Publishing.

- OECD. (2011b). *Tuning-AHELO conceptual framework of expected/desired learning outcomes in engineering* (OECD Education Working Papers, No.60). Paris: OECD Publishing.
- OECD. (2012a, January). *Engineering assessment framework. AHELO feasibility study*. Retrieved November 26, 2016, from [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=edu/imhe/ahelo/gne\(2011\)19/ANN5/FINAL&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=edu/imhe/ahelo/gne(2011)19/ANN5/FINAL&doclanguage=en)
- OECD. (2012b, April). *Economics assessment framework. AHELO feasibility study* (April 2012). Retrieved November 21, 2016, from [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=edu/imhe/ahelo/gne\(2011\)19/ANN3/FINAL&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=edu/imhe/ahelo/gne(2011)19/ANN3/FINAL&doclanguage=en);
- OECD. (2014). *Competency framework*. Retrieved October 17, 2016, from https://www.oecd.org/careers/competency_framework_en.pdf
- Rees, C., Forbes, P., & Kubler, B. (2006). *Student employability profiles. A guide for higher education practitioners*, York. https://www.heacademy.ac.uk/system/files/student_employability_profiles_apr07.pdf
- Sastre-Fullana, P., De Pedro-Gómez, J. E., Bennasar-Veny, M., & Morales-Asencio, J. M. (2014, December). Competency frameworks for advanced practice nursing: A literature review. *International Nursing Review*, 61(4), 534–542.
- Torney Puta, J., et al. (2015). *Assessing civic competency and engagement. Research background, Frameworks, and directions for next-generation assessment. Research Report*, ETS Publication 2015. Retrieved October 16, 2016, from http://www.ets.org/research/policy_research_reports/publications/report/2015/jvdz
- Tremblay, K., Lalancette, D., & Roseveare, D. (2012–2013). *Assessment of higher education learning outcomes AHELO*. Volumes 1–3, Paris.
- Tuning Central Asia Higher Education Area (TuCAHEA) Tempus project. Retrieved October 16, 2016 from <http://www.tucahea.org>
- Tuning Educational Structures in Europe website. <http://www.unideusto.org/tuningeu/publications/subject-area-brochures.html>
- University of Nottingham. *Teaching and learning. Assessment framework*. Retrieved November 21, 2016 from <https://www.nottingham.ac.uk/teaching/assessmentfeedback/assessmentframework.aspx>
- Wagenaar, R. (2013, November). Columbus' Egg? Qualifications frameworks, sectoral profiles and degree programme profiles in higher education. *Tuning Journal for Higher Education*, 1(1), 71–103.
- White House, *Education. Knowledge and skills for the jobs of the future. Higher education*. Retrieved November 26, 2016., from <https://www.whitehouse.gov/issues/education/higher-education>
- Yorke, M. (2006). *Employability in higher education: What it is – What it is not?* Learning & Employability. Series One. York, 2006. [http://www.employability.ed.ac.uk/documents/Staff/HEA-Employability_in_HE\(Is,IsNot\).pdf](http://www.employability.ed.ac.uk/documents/Staff/HEA-Employability_in_HE(Is,IsNot).pdf)

Part III
Generic Student Learning Outcomes –
Cross-National Comparative Approaches

Chapter 10

International Performance Assessment of Learning in Higher Education (iPAL): Research and Development



Richard J. Shavelson, Olga Zlatkin-Troitschanskaia, and Julián P. Mariño

Abstract Educators and policy makers now recognize how important and necessary it is to assess student learning outcomes (SLOs) in higher education. The question has shifted from whether such outcomes should be measured to how they should be measured. Today SLOs are typically assessed by student self-reports of learning or with multiple-choice and short-answer tests. Each of these methods has its strengths and limitations; each one provides insights into the nature of teaching and learning. An alternative approach is the assessment of performance using “criterion” tasks that are drawn from real-world situations in which students are being educated, both within and across academic or professional domains. The international Performance Assessment of Learning (iPAL) project, described herein, consolidates previous research and moves to the next generation of performance assessments for local, national, and international use. iPAL, a voluntary collaborative of scholars and practitioners, seeks to develop, research, and use performance assessments of college students’ twenty-first-century skills (e.g., critical thinking, written communication, quantitative literacy, civic competency and engagement, intercultural perspective-taking) for both instructional improvement and accountability purposes.

R. J. Shavelson (✉)
Stanford University, Stanford, CA, USA
e-mail: richs@stanford.edu

O. Zlatkin-Troitschanskaia
Johannes Gutenberg University, Mainz, Germany
e-mail: lstroitschanskaia@uni-mainz.de

J. P. Mariño
Institute for the Promotion of Higher Education (ICFES), Bogota, Colombia
e-mail: Jp.marino@uniandes.edu.co

10.1 Introduction

The demand to measure higher education learning outcomes has gained worldwide momentum. This is due to both the internationalization and harmonization trends, associated with the globalization of the job market, as well as a very rapid expansion of higher education in developing countries. While several approaches to measuring higher education learning outcomes have been developed over the last decade (see an overview in Zlatkin-Troitschanskaia et al. 2015, 2017a, b), including self-report surveys of learning and multiple-choice and short-answer tests, there are currently very few assessments that focus on direct measures in the form of performance assessments. Perhaps the most famous example is the OECD's project, the Assessment of Higher Education Learning Outcomes (AHELO). This international study examined the feasibility of validly measuring student learning outcomes in higher education internationally and focused among other things on assessments for measuring performance-oriented generic skills (OECD 2012). The AHELO study illustrates both the advances and the challenges that arise when attempting to develop valid and reliable performance assessments.

The international Performance Assessment of Learning (iPAL) project seeks to consolidate previous research and move to the next generation for use locally, nationally, and internationally. It seeks to build a voluntary collaborative of researchers, measurement specialists, and higher education practitioners and supporters with the goal of developing, researching, and using performance-based assessments of learning (PALs) designed to tap college students' twenty-first-century skills for formative and summative purposes.

Based on the AHELO experience with the Collegiate Learning Assessment (OECD 2012, 2013a; www.cae.org), iPAL recognizes that providing performance tasks based on situations drawn from a single national context is limiting (e.g., Zlatkin-Troitschanskaia et al., Chap. 12 in this volume). Tasks need to be developed from multiple national contexts and vetted for their applicability across participating nations and contexts. Hence, forming a collaborative of nations/contexts with representatives from at least Europe, the Americas, and Asia is the vision of the iPAL project.

While drawing a distinction between different varieties of direct measures of learning is somewhat arbitrary (Fu et al. 2016; Shavelson et al. 2017a, b), the focus of iPAL is on measuring the so-called generic twenty-first-century skills. Such skills include critical thinking, analytic reasoning, problem-solving, perspective-taking, and communicating on (at least) some assessment tasks that simulate as closely as possible real-life decision-making and judgment situations (e.g., Shavelson 2012, 2013a, b). These skills are required of all students and graduates in higher education regardless of their field of study and are also viewed as increasingly important by employers and other stakeholders (see also Alexander, Chap. 3 in this volume). Performance assessments are complex tasks that attempt to simulate reality as closely as possible and require test takers to make and justify their decisions and judgments using evidence in the simulated situations (e.g., Shavelson et al. 2015).

Due to their high level of authenticity, these types of tasks are particularly attractive. However, test developers and researchers are faced with a number of conceptual and methodological challenges when creating them (e.g., Shavelson et al. 2015; see also Oser et al., Chap. 7 in this volume). In iPAL, these challenges will be systematically addressed conceptually and methodologically.

Performance assessments that provide valid and reliable measures of learning during and after a student's course of study, and when he or she enters the job market, are relevant for various stakeholders. At the end of a student's course of study, the assessment should provide evidence as to just how competent the student is in the learning outcomes – knowledge, skills, and dispositions in life outside the academy. When those results are contrasted with entry measures and aggregated to produce higher education program measures of skills' development, they can provide valuable evidence of program effectiveness (Shavelson et al. 2016).

Performance assessment tasks can also be used instructionally to teach and assist students in developing the assessed skills and provide corrective feedback on their skill acquisition. This formative assessment property makes this project especially interesting to higher education institutions (HEIs). Given findings from the program Modeling and Measuring Competencies in Higher Education (KoKoHs) (Zlatkin-Troitschanskaia et al. 2016) and the Program for the International Assessment of Adult Competencies (PIAAC) (see OECD 2013b) that higher education graduates seriously lack these skills, the formative function of iPAL seems urgently needed in order to develop appropriate and effective curricular and instructional designs to foster such skills (for curriculum-instruction-assessment triad, see Pellegrino et al. 2001; see also Shavelson 2017). The iPAL project aims to provide HEIs with a high-quality technical framework to address what appears today, in many cases, to be a major challenge: finding a means to foster generic skills and accounting for it. Finally, the near-term outcomes of iPAL might be applicable to performance assessment in specific disciplines and professions (e.g., chemistry, economics, education, medicine, engineering); for now, iPAL focuses on generic skills.

The research focus and goals of the iPAL project as well as the conceptual and methodological background and assessment framework presented in this paper are based on existing knowledge and research. The paper takes a closer look at specific challenges posed by developing performance assessments, in particular for the international assessment of these complex skills. Dealing with these demands and tasks determine the next steps of the iPAL project.

10.2 Existing Knowledge and Research

While the field of assessing learning outcomes in higher education has received increasing attention (Zlatkin-Troitschanskaia et al. 2015, 2017a, b), currently very few projects and assessments focus on measuring students' *performance* on concrete, real-world tasks demanding generic skills. iPAL is based on the previous knowledge and research, most notably from AHELO (OECD 2012, 2013a) and

more specifically with AHELO's experience gained from two Collegiate Learning Assessment (CLA, Shavelson 2010, 2012, 2013a, b) performance tasks adapted and implemented in nine countries.¹ The evidence gained from AHELO provides vivid "lessons" for iPAL – relevant both to the challenges posed by developing these types of tasks as well as when implementing them in higher education both nationally and internationally (OECD 2013a, b).

Building on its CLA, the Council for Aid to Education (CAE) launched the next generation, the CLA+. The revision introduced shorter performance tasks and multiple-choice items so as to produce individual student scores (Zahner 2013). CLA+ is available internationally (Wolf et al. 2014). It has been used not only in the United States (USA) but also adapted and used in Italy and the United Kingdom (UK) (Zahner and Ciolfi, Chap. 11 in this volume). More specifically, this computer-delivered assessment consists of a performance task, where students are confronted with a complex scenario; they are presented with a collection of documents with additional information and data to help them evaluate the case and have to decide on a course of action. The task has an open-ended response format and is complemented by 25 selected-response questions. According to CAE (2013), the performance tasks (PT) measure the following constructs or dimensions:

- Problem-solving and analysis
- Writing effectiveness
- Writing mechanics

With an additional 25 *selected-response questions (SRQ)*, the following student abilities will be measured:

- Reasoning scientifically and quantitatively
- Reading critically and evaluatively
- Critiquing an argument

In Germany, the KoKoHs research team adapted and conducted validation studies with the CLA+ to decide whether to implement it in German higher education (Zlatkin-Troitschanskaia et al., Chap. 12 in this volume). The work consisted of adapting and validating two performance tasks and the accompanying selected-response questions. To this end, several expert workshops and cognitive interviews with students were conducted. Furthermore, German university lecturers evaluated the CLA+'s selected-response questions via an online rating survey (for more details, see Zlatkin-Troitschanskaia et al., Chap. 12 in this volume).

Alongside the projects in the USA and Europe, the importance of measuring students' learning outcomes in higher education is also increasing in Central and South America, where the focus lies particularly on generic skills.² On top of the

¹ Colombia, Egypt, Finland, Korea, Kuwait, Mexico, Norway, the Slovak Republic, and the USA (Connecticut, Missouri, Pennsylvania).

² For the assessment of discipline-specific skills, many different tests and assessments exist in various countries, for example, ETS', Major Field Tests (MFTs) in the USA, and *Exámenes Generales para el Egreso de Licenciatura (EGEL)* by Ceneval in Mexico and KoKoHs in Germany and Austria (see an overview in Zlatkin-Troitschanskaia et al. 2016).

work being carried out in Mexico and Brazil (see an overview in Zlatkin-Troitschanskaia et al. 2015), Colombia has implemented a comprehensive state assessment system for higher education (Shavelson et al. 2016), which includes tests that measure the following generic skills:

- Critical reading
- Quantitative reasoning
- Citizenship
- Written communication

The Education Testing Service (ETS) has developed several assessments of generic skills; among their most advanced tests are the *HEIghTen* tests (cf. Liu et al., Chap. 13 in this volume). They are computer-based tests with closed-ended items in a multiple-choice format. These assessments seek to measure the following generic skills (ETS 2017):

- Critical thinking
- Written communication
- Quantitative literacy
- Civic competency and engagement
- Intercultural competency and diversity

The “critical thinking,” “quantitative literacy,” and “written communication” tests have been developed and validated (e.g., Liu et al. 2016). The others still remain to be developed and tested (ETS 2017).

Apart from higher education, there are many other tests and assessments of learning worldwide (see an overview in Zlatkin-Troitschanskaia et al. 2017a, b). However, so far, only very few of them follow a performance-oriented approach. One such exception is the case for tests to assess professional expertise in the field of vocational training and education as developed in Germany and Switzerland (e.g., Achtenhagen and Winther 2014; Holtsch et al. 2016). In most other learning outcome assessment projects in K-12 education such as OECD’s Programme for International Student Assessment (PISA) and its Program for International Assessment of Adult Competencies (PIAAC), generic skills were not measured using performance assessments.

When developing a conceptual framework for performance assessment of generic skills, some studies focus on conceptually defining the specific competencies and skills that can provide a useful foundation for building an assessment framework. Lai and Viering (2012) (see Table 10.1) and Pellegrino and Hilton (2012) provide examples (see also, e.g., Strijbos et al. 2015).

Table 10.1 Cross-mapping of individual twenty-first-century skills (Lai and Viering 2012)

21CS	P 21 framework subskills	NRC framework subskills	ATC21
Critical thinking	Critical thinking	Critical thinking	Critical thinking, problem-solving, and decision-making
Communication and collaboration	Communication and collaboration	Complex communication and teamwork	Communication and collaboration
Creativity and innovation	Creativity and innovation	Nonroutine problem-solving	Creativity and innovation
Self-regulation and metacognition	Initiative, self-direction	Self-management and self-regulation	Metacognition and learning to learn
Social and cultural competence	Social and cross-cultural skills	Social skills, cultural sensitivity, and dealing with diversity	Local and global citizenship and personal and social responsibility
Flexibility/adaptability	Flexibility and adaptability	Adaptability	NA
Information and technological literacy	Information literacy, media literacy, and information and communications technology literacy	NA	Information literacy and information and communications technology literacy

10.3 Research Focus and Objectives

The iPAL project aims to bring together the best existing expertise from different countries, projects, and initiatives in order to enable cutting-edge fundamental research and practical implementation of performance assessments and (possibly) corresponding teaching-and-learning tools, based on the latest and most innovative state of technology and scholarship.

The research collaborative aims to develop reliable and valid performance assessments of twenty-first-century skills that can be used by higher education institutions nationally and cross-nationally to measure learning outcomes. The focus lies on generic skills that college graduates are expected to develop in order to become engaged citizens of the world. Such skills involve knowledge of content as well as skills such as quantitative reasoning, critical literacy, and written and oral communication that college graduates can draw upon to address life's everyday judgments, decisions, and challenges; they do not include in-depth domain-specific or professional knowledge. The iPAL project focuses on generic skills in part because of their importance to lifelong learning, in part for the wide applicability of such assessments across disciplines and schools in a university, and in part to make the task manageable.

Building on the existing expertise and previous work (see Sect. 10.2), the aim is to develop assessments that focus on generic twenty-first-century skills and that incorporate new research on performance tasks, rational thought, and item formats that integrate innovative media in “real-life” contexts with high fidelity. The goal is to achieve reliable scores for individual test takers. These assessments are what might be thought of as the next generation of performance assessments.

On the one hand, the research goals refer to *test development*. We intend to create and analyze task templates for different generic competencies (see the rows in Table 10.2) and see to what extent they are transferable and adaptable across the different possible tasks’ topics, such as “health” or “arts” (see the columns in Table 10.2). Despite the tasks’ differing topics and contexts (e.g., sports or economics), they are designed to measure the same skills, for example, perspective-taking, and it is therefore important to analyze to what extent they empirically measure generic skills or perhaps even domain-specific abilities. This will then show how generic skills and domain-specific skills correlate with one another. This and other questions will be examined on the basis of a broad validity approach and in accordance with the standards by American Educational Research Association (AERA), Association for Educational Assessment (AEA), and National Council on Measurement in Education (NCME) (2014), which provide a validation framework for issues regarding construct validity, such as correlations with student learning success in a nomological network.

The other goals focus on *test administration and implementation* in higher education and test use in teaching and formative assessment practice. We are concerned about such issues as curricular sensitivity and instructional validity as well as in-depth analyses of individual student and task interaction. This includes questions of cognitive and non-cognitive processes and mental operations while working on test tasks (e.g., Brückner and Zlatkin-Troitschanskaia, Chap. 6 in this volume). In CogLabs (e.g., Leighton 2017), the quality of decision-making can be examined closely, with focus on factors such as reflective vs. intuitive task solving, the influence of (domain-specific) expertise, and test motivation. These analyses, including experimental studies with a pre-post-design, can provide an empirical basis for important implications when designing new curricula, instructions, assessments, and feedback in higher education practice. They serve as the basis of rigorous experimentation.

Other goals become apparent when *developing and implementing assessments in international studies*. Besides substantial challenges dealing with test translation, adaptation, and validation across countries, there are also other specific challenges depending on different types of assessment tasks and validation procedures. Not only language-cultural influences must be examined but also the effects of different task formats and parallel test versions.

10.4 Conceptual and Methodological Background: Assessment Framework

10.4.1 Holistic Approach

The iPAL project is not based on the usual assessment framework. Usually such frameworks divide the construct – twenty-first-century skills – into component parts, such as critical thinking, problem-solving, perspective-taking, and communicating, and then each is divided further into its component parts, and a “uni-dimensional” measure is developed for each subcomponent. Once this is done, the internal structure of the assessment is examined, and reliability and validity evidence is presented.

Instead, iPAL takes a *holistic approach* to the development of an assessment of twenty-first-century skills. The whole is viewed as greater than the sum of its parts. Real-world situations demanding the application of these skills do not come nicely divided into component parts. Rather, more likely, as Snow has demonstrated, subsets of these skills are sequenced over the course of addressing challenges (Corno et al. 2002). More specifically, “by aptitudes, Snow meant all ... [those] characteristics (e.g., experience, ability, knowledge, motivation, and regulatory processes) that an individual brings to and cobbles together to perform in a particular situation. He called this situation-elicited set of aptitudes an “aptitude complex.” (Snow 1996) Over time, individuals might attend to different aspects of the situation (test) and bring somewhat different aptitude complexes to bear. That is, these aptitude complexes were viewed as dynamic – they changed in relation to changes in the task environment, changes that often are brought about by an individuals’ own actions as they move through a task” (Shavelson et al. 2002, p. 79). Consequently, and in accordance with Snow (1996), iPAL proposes to sample real-world events (plenty are provided in the morning newspaper) and adapt them in an assessment framework that provides definition, organization, and a means of scoring responses to tasks with multiple completion paths.

The iPAL project recognizes the various needs of HEIs that may include, for example, reliable assessments of critical thinking or quantitative reasoning (e.g., Alexander, Chap. 3 in this volume). In this case, the performance assessment would include multiple performance and selected-response tasks that tap various aspects of critical thinking or quantitative reasoning (e.g., Shavelson et al. 2017a, b).

This holistic approach is also embodied in what has been called a criterion-sampling approach to measurement (McClelland 1973). This approach too assumes that the whole is greater than the sum of its parts and that complex tasks require an integration of abilities that cannot be captured when divided into and measured as individual components. The criterion-sampling notion is straightforward: If a researcher wants to know what a person knows and can do, they should sample tasks from the domain in which that person is to act, observe her performance, and infer competence and learning. For example, a person’s ability to drive a car should not be assessed simply by a multiple-choice test, which would

be suited only to assess whether the person knows the laws governing driving a car. To assess actual driving ability, one would also administer a practical driving test with a sample of tasks from the general driving domain such as starting the car, pulling into traffic, turning right and left in traffic, backing up, and parking. Based on this sample of performance, it is possible to draw more generally valid inferences about driving performance.

We propose sampling tasks and collecting students' "operant responses" as well as "respondent" responses (McClelland 1973). Operant responses are student-generated responses that are modified with feedback as the task is carried out; respondent responses are selected responses (usually to multiple-choice questions). These responses are to parallel those expected on real-world tasks and activities that are organized and developed in such a way as to test twenty-first-century skills.

In what follows, we provide a sketch of the assessment framework including working definitions of the generic skills and the tasks created to assess them. The tasks can be weighted in a way that emphasizes performance tasks of problem-solving and others on, say, quantitative reasoning. Nevertheless, all have the same underlying critical thinking dimensions.

10.4.2 Construct Definition

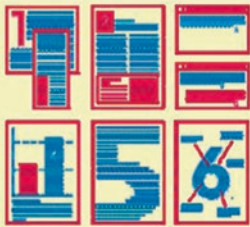
The overarching construct underlying PALs is the competence of citizens to think critically, solve problems, take the perspective of others, and communicate clearly their ideas, beliefs, analyses, etc. (see Table 10.1) when confronted with everyday complex life situations. These are called twenty-first-century skills or "generic" skills for lack of a better name and to avoid the jargon of twenty-first-century skills.

Initially, five categories of generic skills for iPAL assessment were identified (see Table 10.2). To a greater or lesser extent, two or more such skills would comprise an assessment. Typically, a real-world event or "problem to be solved" would be presented along with information more or less relevant to the event or problem. The problem might be similar to that in Fig. 10.1, for example, one that requires critical thinking to combine both pieces of data, and some elementary quantitative reasoning to solve. In another case, the problem might be visual spatial, for example, in creating an art exhibition that involves a tension between engineering progress and negative impacts on the environment. And at other times it might be verbal, in which varying sides to a proposed civic project – where to situate a prominent movie mogul's museum if at all – are aired and an understanding of these various perspectives is needed to make progress.

Three elements are introduced into the problem or event that are likely to evoke critical thinking: (1) the reliability of the information source, (2) the validity of the information for the particular problem or event at hand, and (3) the information's susceptibility to judgmental errors when thinking too quickly (cf. also Alexander, Chap. 3 in this volume).

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of DynaTech's sales force, recommended that DynaTech buy a small private plane (a SwiftAir 235) that she and other members of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation:

- 1: Newspaper articles about the accident
- 2: Federal Accident Report on in-flight breakups in single engine planes
- 3: Pat's e-mail to you & Sally's e-mail to Pat
- 4: Charts on SwiftAir's performance characteristics
- 5: Amateur Pilot article comparing SwiftAir 235 to similar planes
- 6: Pictures and description of SwiftAir Models 180 and 235



Please prepare a memo that addresses several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into account, and your overall recommendation about whether or not DynaTech should purchase the plane.

Fig. 10.1 Airplane task (Shavelson 2013a, p. 78)

10.4.3 Task Universe

The universe of tasks demanding generic skills comprises the myriad everyday complex life situations. The iPAL samples such situations for inclusion in performance tasks and more traditional items (e.g., multiple-choice; see below). A prime source of situations may be found easily in newspapers (e.g., politics, environment, sports, business, fashion, arts, and science; see Table 10.2). The airplane task described below (see Fig. 10.1), for example, was inspired by the report of an aircraft crash at the Van Nuys Airport in Southern California.

These tasks are complex often without a clear path toward solution, decision, or action. Rather there are trade-offs. They admit to more than one solution although when incorporated into an assessment, they have better and worse solutions, decisions, actions, etc. The tasks are compelling in the sense they represent current everyday challenges that test takers face or might be expected to face as college graduates and citizens more generally.

10.4.4 Elements of Critical Thinking

Critical thinking is conceived as the process of conceptualizing, analyzing or synthesizing, and evaluating and applying information to solve a problem, decide on a course of action, find an answer to a given question, or reach a conclusion. Assessment tasks are developed to include certain elements that invite students to think critically. These elements are (e.g., Shavelson 2010):

Information Source Sampling

Materials such as newspaper articles, YouTube videos, and government reports are sampled from real-world domains (see above). The information provided may be manipulated to be either:

- (a) *Reliable* or trustworthy such as the Federal Aviation Administration (FAA) report in the airplane task³; in contrast, an amateur aviator's opinion article would be considered to be less or unreliable.
- (b) *Valid* or directly relevant to the issue at hand (FAA report) or tangential or unrelated to the task (photos of the SwiftAir 135 and 235).

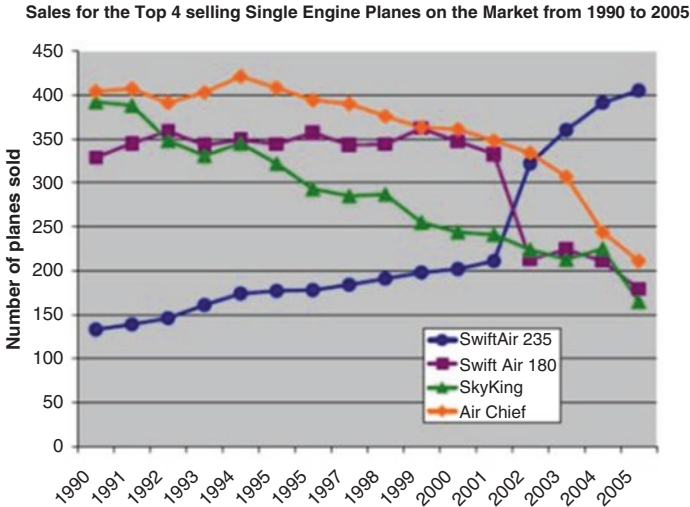
Judgmental and Decision Heuristic Sampling

In using information to make judgments and decisions, people often take shortcuts or use heuristics to make judgments or reach a decision (e.g., Krolak-Schwerdt et al., Chap. 5 in this volume). The work of Tversky and Kahneman (1974) opened up a field that has become known as rational thought (e.g., Kahneman 2011; Stanovich 2009). These heuristics are normally applicable in the real world, where quick judgments or decisions must be made and deliberative thought might be dangerous (e.g., get out of the crosswalk because the car isn't going to stop). Heuristics have been quite useful throughout evolution. However, they can interfere with rationality – critical thinking or problem-solving – when the situation is important enough to demand a rational decision (e.g., buying a house). In this case, deliberative thought is needed to simulate alternatives and their consequences. Since Tversky and Kahneman's initial research, the list of judgmental and decision-making heuristics has exploded (e.g., Stanovich 2016) and can be easily researched on the Internet. Consequently, irrational (when the situation demands otherwise) thinking heuristics are built into performance tasks or might be assessed in stand-alone multiple-choice questions.

The airplane task (see Fig. 10.1) uses one of those heuristics where baseline conditions (number of aircraft sold) are ignored and unadjusted data are used to make decisions. From Fig. 10.2, leaving sales aside, one would conclude (problematically) that the SwiftAir 235 is, indeed, more accident prone than its competitors.

There are many other heuristics that can be incorporated into the assessment tasks that simulate, with high fidelity, everyday events. Moreover, the aim is to create a separate selected-response portion of the PAL that probes students' ability to resist "fast thinking" and slow down to "simulate" alternative courses of action and their alternatives. Finally, the framework includes incorporating the capacity to take others' point of view in assessing problem solutions, alternative courses of action, and the like.

³Note that a government report in the USA, such as the Federal Aviation Reports on aircraft accidents, are considered to be highly reliable. However, in other countries, government reports are treated with great suspicion and not considered to be reliable. Hence the challenge in developing tasks that cross boundaries.



Note: Only companies with more than 10% of the market share are included in the above figure.

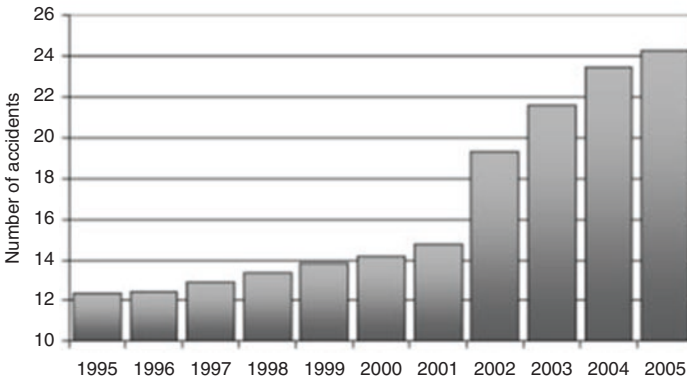
12-15-06

Pat, I plotted some data that I found on the FAA’s website. It appears that there was a significant increase in the number of SwiftAir 235 accidents after the company switched from the strut braced wing to the cantilevered wing. Based on this it looks like the new wing might not be so safe after all.

Harrison

Number of Accidents by the Top 4 Selling Single Engine Planes

SwiftAir 235 and 235-C accidents



Note: This chart includes all types of accidents (in flight and on the ground) that resulted in injury requiring physicians’ attention or at least \$500 in damage.

Fig. 10.2 Quantitative information provided in DynaTech performance task (Shavelson 2008, p. 36)

10.4.5 *Elements of Communicating*

The ability to communicate clearly, concisely, accurately, and compellingly is part of our conception of generic skills. The communication might be in writing (e.g., a memo to the president of a company or an op-ed piece), orally with visuals, or both (e.g., PowerPoint presentation with notes). Such a communication would:

- Use reliable information and avoid less-than-reliable information
- Use relevant information and avoid peripheral information
- Avoid judgmental and decision-making “traps”
- Consider alternative courses of action to the one proposed and indicate why the recommendation is given
- Use concise compelling arguments from evidence to conclusions to rhetorically establish a position, decision, course of action, or recommendation

10.4.6 *Example Application of Assessment Framework*

The envisioned assessments are what might be thought of as the next generation of performance assessments, moving beyond, for example, the work of the AHELO with the Collegiate Learning Assessment (CLA) (www.cae.org). As an example of what is envisioned, consider the assessment task, “DynaTech,” drawn from an early CLA (the task shall be called “airplane” for ease; see Fig. 10.1). The task asks students to advise the president of a company who is about to purchase an aircraft for business purposes (Shavelson 2013a). The aircraft that is about to be purchased has had an accident, and the question arises as to whether the airplane is safe. Students are asked to use a variety of information sources (e.g., newspaper articles, Federal Aviation Agency report, an opinion piece by an amateur aviator) to determine whether the aircraft is accident prone and whether the company should move forward with the purchase.

In one of the subtasks, students need to decide whether the aircraft in question, the SwiftAir 235, is indeed accident prone. In the (reliable and valid) information provided, the student sees two panels of data (see Fig. 10.2). One panel provides information on aircraft sales, and the second panel provides data on accidents of the SwiftAir 235 and competitor aircraft. With this information, the student is in a position to make a determination as to whether the SwiftAir 235 is accident prone. Note that if the student focuses on the panel showing the number of accidents – information directly related to the accident-prone question – the conclusion is that the aircraft is accident prone. However, if the student stops a minute and considers the sales data in conjunction with the number of accidents, the accident *rate* rather than the number of accidents becomes available, with the conclusion that the SwiftAir 235’s accident rate turns out to be the lowest. In combining both pieces of information, the task invites students to use a fast-thinking heuristic (Kahneman

2011) and avoid the trap of not considering baseline information or to think more slowly and draw a more justified conclusion (cf. Stanovich 2009, 2016; Alexander, Chap. 3 in this volume).

Building on AHELO and other work (e.g., Shavelson 2013a, b), the aim is to create an assessment that focuses on generic twenty-first-century skills that incorporates new research on rational thought, that goes beyond the current item formats to incorporate, for example, video and spreadsheets, and that produces reliable scores for individual test takers.

Moreover, the assessment framework should specify in detail the following characteristics (among others, as work proceeds).

10.4.7 *Task Formats*

PALs will be delivered on a computer platform and in many cases over the Internet (depending on security). Computers provide substantial leeway both in delivering tasks and in their fidelity to the real world they are intended to emulate. The task-format decision is driven first and foremost by its fidelity to the criterion situation being simulated. This said, cost and safety are also important considerations, and they too must be incorporated into the selection of formats.

There is a possibility for multiple formats. Some formats will be *open-ended*, and students will construct answers of varying length in response to a prompt inviting them to make a judgment or decision, to recommend a course of action, to solve a problem, and so on. At least one subtask will be of sufficient length to evaluate students' writing as to the (1) evidence presented from information provided to justify a decision or recommend a course of action and (2) clarity and force of argument presented.

Selected-response (e.g., multiple-choice) formats can be used to probe critical reading of, for instance, documents provided in the task, quantitative reasoning with graphs or tables (etc.) provided, or rational thinking with stand-alone prompts (Stanovich 2016). Still other formats can be brief, self-contained tasks with either short constructed responses or multiple-choice questions (see also Oser et al., Chap. 7 in this volume).

10.4.8 *Scoring*

For the extended constructed responses, analytic (dimensional) scoring rubrics will be developed based on the construct definition. This means that the rubrics can take into account the test-takers' use of reliable and unreliable and valid and invalid information as well as their reflection and avoidance of heuristics that lead to errors in judgment and decision-making. The rubrics can examine the use of such

information in justifying decisions, problem solution, and/or recommendations for action. Moreover, they can evaluate argumentation, the use of evidence to support claims, and clarity of communication.

10.4.9 Assessment Delivery

The medium of delivery, as noted above, may vary widely, taking advantage of computer affordances. For example, a spreadsheet might be used for calculations, simulations might be used for modeling alternatives, and PowerPoint might be used for presentation and justification of, say, recommendations. An intranet containing reliable and unreliable documents, relevant or irrelevant documents, etc., might be used to examine students' capacity to search and bring evidence to bear on a problem. Audio may be used to enhance the fidelity of the simulated situation or to collect students' verbal "presentations" of findings. In the final analysis, the technology is subservient to the construct measured, not vice versa. However, the technology provides a means of increasing simulation fidelity over what is possible with pencil and paper.

10.5 Further Research Perspectives and Demands

10.5.1 Challenges in Developing Performance Assessments

The project as a whole is challenging. Performance assessment is used to measure performance in education, work, and everyday life. Such assessment presents an activity or set of activities that requires test takers, individually or in groups, to generate products or performances in response to a complex task. These products or performances provide observable or inferable evidence of the test taker's knowledge, skills, abilities, and higher-order thinking skills in an academic content domain, in a professional discipline, or on the job.

The psychometric challenges of performance assessment are often treated as the proverbial elephant in the room. The challenge, simply put, is that performance assessment involves complex, lifelike tasks and parallel real-life responses that can be intricate, lengthy, and limited in number due to time and cost. Standard psychometric models were developed for multiple-choice assessments, with many discrete test items that are scored dichotomously and that are organized into tests designed to measure one clearly defined construct. Performance assessment is complex to model and implement and requires thorough testing and examination to confidently associate test-taker performance with a score or a performance category (see, e.g., Oser et al., Chap. 7 in this volume).

Shavelson et al. (2015, pp. 97–98) detail and summarize such challenges. Some of the psychometric challenges associated with modeling performance assessment are:

- **Limited number of observations:** Psychometric models work best when there are many of the same kind of observations of the same construct (e.g., many questions to assess reading comprehension). The time and cost associated with performance assessment put a practical limitation on the number of observations that are possible.
- **Complex and varied score scales:** Performance assessments are not generally scored simply as either right or wrong. They might be scored using a rubric, or multiple rubrics, on scales that range from 0 to 3 or 1 to 6 or any other variant (percentages, error rate). They may also be scored using more unusual scales, such as the time required for a test taker to respond or some other process indicators. Further, the same performance assessment may result in multiple scores of different types.
- **Human influence (raters):** Performance assessments are often scored by human judgment. That is, raters are trained to read or observe student work and evaluate it based on the defined scoring criteria (e.g., rubrics). While a high level of training and monitoring greatly helps to ensure rater accuracy, rater variation can introduce measurement error.
- **Human influence (group members):** Human influence can be particularly bothersome when the assessment is conducted in the context of groups. A student's performance on a group work skill (e.g., the ability to consider the ideas of others) is likely to be influenced by the behavior of the others in the group.
- **Connectedness:** The tools that psychometricians use to convert test-taker performance to a score or category work best when various test questions/activities are unconnected (i.e., they satisfy the assumption of local independence). A performance assessment typically includes a set of activities, products, and item types that are designed to be connected. A complex performance assessment task, for example, that requires a medical student to collect information and make a diagnosis may result in multiple scores based on many decisions or processes, but the scores would all be related to, for example, the same patient situation.
- **Dimensionality:** Most psychometric models work best when an assessment measures one construct at a time (i.e., assumption of unidimensionality), so that the interpretation of the resulting score or performance category is clear. A performance assessment that requires a mathematics student to solve a complex multistep problem and then write about that process measures the student's ability to demonstrate multiple skills and therefore could confuse the interpretation of the results.

10.5.2 *Challenges in Developing Assessment Tasks for an International Study*

Results from OECD's AHELO have indicated that international comparative assessments are possible but very challenging in terms of conceptualization, methodology, and harmonization. Most existing assessments are locally developed achievement assessments of national scope (Zlatkin-Troitschanskaia et al. 2017a, b), while reliable and valid international assessments of student learning are scarce. The iPAL project strives to combine the best of both worlds by offering task blueprints (and further design recommendations) for the development of locally valid tasks while also making the best tasks available for international adaptation to form an internationally comparable task pool.

Challenges in international assessment result from greater organizational demands including the need for additional coordination and consensus building as well as from needs for harmonization or explicit differentiation of the assessments across a wider range of educational frameworks, learning objectives, curricula, teaching-and-learning cultures, conceptualizations, assessment purposes, stakeholder interests, etc. To obtain internationally comparable results, work done by different national teams in test development and adaptation and administration has to be closely monitored and attuned to ensure harmonization (Test Adaptations Guidelines (TAGs) by the International Test Commission (ITC) (2005). Furthermore, additional comparability analyses are necessary to establish measurement invariance across comparable groups in different countries (see also Hambleton and Zenisky 2010).

The most promising approach to systematically generate functionally equivalent international assessment tasks is to integrate adaptation and test development, aiming for top-down comparability and harmonization at every step from construct definition, assessment framework design, definition of the target population, task operationalization (including definition of construct-relevant and construct-irrelevant parts) to create so-termed *conceptual task shells* (Solano-Flores et al. 2001) for generating highly specific tasks with similar structures and appearances that are adaptable across nations, as well as for subsequent translation and adaptation, and quality assurance in revisions. To ensure comparability, administration and interpretations need to be comparable, as well, which requires implementation of similar processes across nations for pretesting, sampling, task administration, technical presentation, incentives, and validity and comparability analyses (e.g., Marion and Pellegrino 2007; Pellegrino et al. 2001; AERA, AEA and NCME 2014).

Psychometric quality criteria need to be confirmed in each country individually based on a comprehensive, if possible coordinated, validity concept that aligns theoretical and empirical evidence from the test scores and the interpretations of these test scores to indicate, most importantly, whether the test score sufficiently represents the targeted construct; for higher education, these are, for example, real-life and/or

job-related requirements (e.g., critical thinking, decision-making, problem-solving; see Table 10.2). Comparable processes and quality benchmarks need to be linked in validity arguments within countries and with the same quality indicators across countries to maximize both local applicability and international comparability.

Careful planning and problem resolution routines are needed to address arising questions, such as what to do if items are not adaptable, if target samples sizes are not reached, if assessments are administered differently, if incentives vary, if quality indicator benchmarks are not met in individual countries, etc. For meaningful interpretations that can help improve education, additional variables need to be assessed and harmonized across participating countries, to enable controlling for educational input and process factors that correlate with test results, for example, including controlling in a valid way for students' individual preconditions, such as the educational path and learning opportunities they have taken as well as sociodemographic variables or belonging to certain groups. Illustrating the importance of assessing such additional variables, international studies in higher education have shown that even with standardized samples across institutions, students' motivation to perform well on tests varies substantially and is the second-best predictor of test performance (the best being students' entry conditions).

Documentation of the uses of tasks, blueprints, and further task development by higher education institutions needs to be prepared, and communication needs to be coordinated internationally. Assumptions underlying test development and the intended uses (e.g., to improve learning outcomes at a higher education institution, to monitor multiple universities) need to be made explicit early on and negotiated in a way across nations to enable a wide enough range of possible uses and interpretations that can be drawn without compromising psychometric quality (e.g., matching tests to specific uses and inferences, avoiding overly broad inference and "function creep," which can also compromise comparability, Koretz 2016).

In the iPAL project, responsibilities of national implementation and coordination and harmonization support can be split between different levels (international) to national and perhaps more local), and the project can draw on experience and methods from international large-scale studies in higher education, the school sector, and comparative survey research (cf. Cross-Cultural Survey Guidelines).⁴

Overall, the iPAL project comprises an ambitious and comprehensive research and development program, which involves multiple milestones and stages. With a view to the existing research demands as well as conceptual and methodological challenges when developing PAL and implementing the assessments in higher education practice, iPAL aims to achieve significant progress in the area of assessing student learning outcomes and, on this basis, promote the acquisition of such outcomes in a systematic manner.

⁴For more details of challenges of international assessment, see also Zlatkin-Troitschanskaia et al. (2015, 2017).

References

- Achtenhagen, F., & Winther, E. (2014). Workplace-based competence measurement: Developing innovative assessment systems for tomorrow's VET programmes. *Journal of Vocational Education & Training*, 66, 281–295. <https://doi.org/10.1080/13636820.2014.916740>.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., & Talbert, J. E. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. New York: Routledge.
- Council for Aid to Education. (2013). Introducing CLA+. *Fostering great critical thinkers*. New York: CAE. http://cae.org/images/uploads/pdf/Introduction_to_CLA_Plus.pdf
- Educational Testing Service (ETS). (2017). *Introducing the HEIghten: Outcomes assessment suite*. <https://www.ets.org/heighten>
- Fu, A. C., Kannan, A., Shavelson, R. J., Peterson, L., & Kurpius, A. (2016). Room for rigor: Designs and methods in informal science education evaluation. *Visitor Studies*, 19(1), 12–38. <https://doi.org/10.1080/10645578.2016.1144025>.
- Hambleton, R. K., & Zenisky, L. (2010). *Translating and adapting tests for cross-cultural assessments*. <https://doi.org/10.1017/CBO9780511779381.004>
- Holtsch, D., Rohr-Mentele, S., Wenger, E., Eberle, F., & Shavelson, R. J. (2016). Challenges of a cross-national computer-based test adaptation. *Empirical Research in Vocational Education and Training*, 8(18), 1–32.
- International Test Commission. (2005). *International Test Commission guidelines for translating and adapting tests*. Retrieved from http://www.intestcom.org/files/guideline_test_adaptation.pdf
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Koretz, D. (2016, April 4). *Measuring postsecondary competencies: Lessons from large-scale K-12 assessments*. Presentation at the KoKoHs conference, Berlin.
- 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.
- Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. Oxford: Oxford University Press.
- Liu, O. L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing critical thinking in higher education: The HEIghten™ approach and preliminary validity evidence. *Assessment & Evaluation in Higher Education*, 41(5), 677–694. <https://doi.org/10.1080/02602938.2016.1168358>.
- Marion, S. F., & Pellegrino, J. (2007). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement Issues and Practice*, 25, 47–57.
- McClelland, D. C. (1973). Testing for competence rather than intelligence. *American Psychologist*, 28, 1–14.
- OECD. (2012). *Assessment of higher education learning outcomes. Feasibility study report: Volume 1 – Design and implementation*. Retrieved from <http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf>
- OECD. (2013a). *Assessment of higher education learning outcomes. AHELO feasibility study report – Volume 2. Data analysis and national experiences*. Paris: OECD.
- OECD. (2013b). *The survey of adult skills: Reader's companion*. OECD Publishing. <https://doi.org/10.1787/9789264204027-en>.
- Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.
- Shavelson, R. J. (2008). Reflections on quantitative reasoning: An assessment perspective. In B. L. Madison & L. A. Steen (Eds.), *Calculation vs. context: Quantitative literacy and its implications for teacher education*. Washington, DC: Mathematical Association of America.
- Shavelson, R. J. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford: Stanford University Press.
- Shavelson, R. J. (2012). Assessing business-planning competence using the collegiate learning assessment as a prototype. *Empirical Research in Vocational Education and Training*, 4, 77–90.
- Shavelson, R. J. (2013a). On an approach to testing and modeling competence. *Educational Psychologist*, 48(2), 73–86.
- Shavelson, R. J. (2013b). An approach to testing and modeling competencies. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education: Tasks and challenges*. Boston: Sense.
- Shavelson, R. J. (2017). Statistical significance and program effect: Rejoinder to “why assessment will never work in many business schools: A call for better utilization of pedagogical research”. *Journal of Management Education*, 41, 1–5.
- Shavelson, R. J., Roeser, R. W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., Schultz, S., Quihuis, G., & Gallagher, L. (2002). Richard E. Snow’s remaking of the concept of aptitude and multidimensional test validity: Introduction to the special issue. *Educational Assessment*, 8(2), 77–100.
- Shavelson, R. J., Davey, T., Ferrara, S., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Princeton: Educational Testing Service.
- Shavelson, R. J., Domingue, B. W., Mariño, J. P., Molina-Mantilla, A., Morales, J. A., & Wiley, E. E. (2016). On the practices and challenges of measuring higher education value added: The case of Colombia. *Assessment and Evaluation in Higher Education*, 41(5), 695–720.
- Shavelson, R. J., Marino, J., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2017a). Reflections on the assessment of quantitative reasoning. In B. L. Madison & L. A. Steen (Eds.), *Calculation vs. context: Quantitative literacy and its implications for teacher education*. Washington, DC: Mathematical Association of America. (in press).
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Marino, J. (2017b). Performance indicators of learning in higher education institutions: Overview of the field. In E. Hazerkorn, H. Coates, & A. Cormick (Eds.), *Research handbook on quality, performance and accountability in higher education*. Edward Elgar. (in press).
- Snow, R. E. (1996). Aptitude development and education. *Psychology, Public Policy, and Law*, 2(3/4), 536–560.
- Solano-Flores, G., Shavelson, R. J., & Schneider, S. A. (2001). Expanding the notion of assessment Shell: From task development tool to instrument for guiding the process of science assessment development. *Revista Electrónica de Investigación Educativa*, 3(1), 33–53.
- Stanovich, K. E. (2009). *What intelligence test miss: The psychology of rational thought*. New Haven: Yale University Press.
- Stanovich, K. E. (2016). The comprehensive assessment of rational thinking. *Educational Psychologist*, 51, 1–12. <https://doi.org/10.1080/00461520.2015.1125787>.
- Strijbos, J., Engels, N., & Struyven, K. (2015). Criteria and standards of generic competences at bachelor degree level: A review study. *Educational Research Review*, 14, 18–32. <https://doi.org/10.1016/j.edurev.2015.01.001>.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Wolf, R., Zahner, D., Kostoris, F., & Benjamin, R. (2014). *A case study of an international performance-based assessment of critical thinking skills*. New York: Council for Aid to Education.

- Zahner, D. (2013). *Reliability and validity of CLA+*. http://cae.org/images/uploads/pdf/Reliability_and_Validity_of_CLA_Plus.pdf
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, *40*(3), 393–411.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Lautenbach, C., & Toepper, M. (2016). Assessment practices in higher education and results of the German research program modeling and measuring competencies in higher education (KoKoHs). *Research & Practice in Assessment*, *11*, 46–54.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Lautenbach, C., Molerov, D., Toepper, M., & Brückner, S. (2017a). *Modeling and measuring competencies in higher education. Approaches to challenges in higher education policy and practice*. Wiesbaden: Springer.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Pant, H. A. (2017b). Assessment of learning outcomes in higher education – International comparisons and perspectives. In C. Secolsky & B. Denison (Eds.), *Handbook on measurement, assessment and evaluation in higher education* (2nd ed.). New York: Routledge.

Chapter 11

International Comparison of a Performance-Based Assessment in Higher Education



Doris Zahner and Alberto Ciolfi

Abstract In late 2012, the Italian National Agency for the Evaluation of the University and Research Systems (ANVUR) and the Council for Aid to Education (CAE) collaborated on two experimental studies of an assessment of generic learning outcomes of tertiary students. The instrument (CLA+ International/TECO) was administered to students graduating from Italian universities. Their results and outcomes were benchmarked against graduating university students in the United States. This chapter presents results from the two studies and discusses implications and future investigations and demonstrates the feasibility of internationally assessing generic learning outcomes.

11.1 Introduction

International assessments in higher education are especially challenging because differences across countries (e.g., educational systems, SES) increase the complexity of testing (Blömeke et al. 2013; Wollack 1997). This becomes even more challenging when using performance-based assessments, which are becoming more prominent in assessment programs (Kahl 2008; Penfield and Lam 2000).

Country participation in international comparative studies such as the Programme for International Student Assessment (PISA) has grown over time. For instance, PISA included 43 participating countries during the first administration in 2000, and that number has grown to over 70 for the 2015 administration (OECD 2016). International comparison of students' achievement at the secondary school level allows countries the opportunity to benchmark their educational system and to identify program strengths and weaknesses in an attempt to enhance instructional

D. Zahner (✉)
Council for Aid to Education (CAE), New York, NY, USA
e-mail: dzahner@cae.org

A. Ciolfi
Agency for the Evaluation of Universities and Research Institutes (ANVUR), Rome, Italy
e-mail: alberto.ciolfi@anvur.it

effectiveness and student learning. While these cross-country comparisons are feasible at the secondary school level, instruments that allow for cross-cultural comparisons at the tertiary school level are much less common (see also Shavelson et al., Chap. 10, in this volume). The measurement of higher-order competencies in higher-education institutions across nations presents challenges due to differences in educational systems, socioeconomic factors, and perceptions as to which constructs should be assessed (e.g., Blömeke et al. 2013).

International academic institutions of higher education are under pressure to enhance the quality of instruction for accountability reasons. In fact, the principal goals of higher education are both academic research and attaining high-quality student learning outcomes (cf. Chap. 3, in this volume). Moreover, the development of higher-order skills could be helpful for the next generation's workforce in order to meet the demands of careers evolving in the twenty-first century. Research suggests that employers seek individuals who are able to think critically and communicate effectively (e.g., Hart Research Associates 2006). In order to meet the demands of today's world, a shift in assessment strategies is necessary to measure the skills now prized in a complex global environment. More specifically, assessments that only foster the recall of factual knowledge have been on the decline, whereas assessments that evoke higher-order cognitive skills, such as analytic and quantitative reasoning, problem-solving, and written communication are on the rise.

CAE's Collegiate Learning Assessment (CLA+) is a performance-based assessment that measures higher-order thinking skills at the tertiary level within the United States and internationally.

The purpose of this chapter is to present two studies that investigate the translation, adaptation, and administration of the CLA+ International. CAE collaborated with ANVUR (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca) to conduct a pilot study in 2012 and a follow-up study in 2015. Rather than viewing and/or treating the assessment in isolation, CAE and ANVUR designed the studies in order to collect evidence for valid cross-national comparisons. The chapter presents the entire translation and adaptation process as well as the results from the two studies.

11.2 Rationale for Generic Skills

Indeed, a college education has never been more necessary for productive participation in society. Employers now seek individuals able to think critically and communicate effectively in order to meet the requirements of the new knowledge economy (e.g., Hart Research Associates 2006; Levy and Murname 2004, see also Chap. 3, in this volume). Therefore, the skills taught in higher education are changing; less emphasis is placed on content-specific knowledge, and more is placed on general higher-order skills such as analytic reasoning and evaluation, problem-solving, and written communication.

Any rigorous improvement project requires continual evaluation in order to measure progress toward goals. Consequently, there is a clear need for standardized assessments such as CLA+ that measure generic skills. Performance assessments like CLA+ not only evaluate whether students are learning higher-order skills required of today's workforce but also spur educational advances in pedagogy. The CLA+ presents students with scenarios that are representative of the types of problems they will encounter in the real world and asks them to generate solutions to these problems. Unlike multiple-choice questions, where students need only to identify the correct answer—limiting the capacity of those questions to measure students' critical-thinking skills—an open-ended assessment such as the CLA+ is able to measure how well students formulate hypotheses, recognize fallacious reasoning, and identify implicit and possibly incorrect assumptions.

Obviously, knowledge and skills specific to academic disciplines are important, but there is a multitude of disciplines, each potentially differing across national contexts and evolving over time. Since different disciplines can require different skills and types of reasoning, it is difficult to establish broad, cross-national benchmarks based on achievement in academic disciplines (cf. Chap. 9, in this volume). The approach of the generic skills strand is to establish benchmarks for student achievement of essential higher-order skills that cut across national contexts and academic disciplines. The development of students' generic skills is central to the missions of modern postsecondary institutions because of growing recognition that these skills fuel innovation and economic growth (Levy and Murnane 2004).

Beginning with participation in the AHELO feasibility study (OECD 2011, 2012a, c, 2013a), CAE has demonstrated that although there are several methodological challenges to measuring student learning outcomes both within and across cultures (Wolf et al. 2015), it is possible to address these challenges and mitigate some of the issues (Wolf and Zahner 2015).

11.3 Method

In 2013, 5853 students from 12 participating Italian institutions completed a translated and adapted version of the CLA+ that included a performance task ("parks") and a set of 20 selected-response questions. In 2015, a new cohort of 6268 students from 23 institutions participated in a second study. The results presented in this chapter reflect the results from both studies.

11.3.1 CLA+ Instrument

CLA+ is a performance-based assessment of critical-thinking and written-communication skills. It consists of two sections, a performance task (PT), which requires students to generate a written response to a given scenario, and

selected-response questions (SRQs).¹ Students have 90 min to complete the two sections of the assessment—60 min for the PT and 30 min for the SRQs.

For the PT, students are given a scenario and asked to make a decision or recommendation after analyzing a document library that contains various sources of information, such as letters, maps, and graphs. They are then expected to write a response to the scenario justifying their decision/recommendation and provide reasons and evidence against the opposing argument(s). The student responses to the PT are measured on three subscales: analysis and problem-solving (APS—identifying, interpreting, evaluating, and synthesizing pertinent information and proposing a solution in terms of how to proceed in case of uncertainty), writing effectiveness (WE—producing an organized and cohesive essay with supporting arguments), and writing mechanics (WM—demonstrating command of written native language). The SRQ section consists of a set of 25 questions that are also document based and designed to measure the same construct as the APS sub-score of the PT. Ten measure scientific and quantitative reasoning (SQR) (e.g., making an inference), ten measure critical reading and evaluation (CRE) (e.g., identifying assumptions), and five measure critiquing arguments (CA) (e.g., detecting logical fallacies). Students are given 60 min to construct a response to the PT and 30 min to respond to SRQs.

For the project, ANVUR rebranded the CLA+ International as TECO (Test sulle competenze di carattere generalista), and references to TECO throughout the chapter are to the translated and adapted Italian version of the CLA+.

11.3.2 *Participants (2013)*

In the design of the translated and adapted version of CLA+ (TECO), ANVUR (2014) established a series of criteria based upon the awareness that the collaboration was a feasibility study with tight deadlines, a limited budget, and voluntary student participation and upon the need to collect as much demographic and contextual data as possible for a more complete understanding of students. Only students, who were qualified to graduate, as defined by their progress through university, were eligible to participate in TECO. Accordingly, the eligible students for TECO 2013 were those enrolled in the third or fourth year of a three-year course or single-cycle master's course who had acquired all the necessary study credits (basic and characterizing).

The demographic summary of the participating students compared to all eligible students is shown in Tables 11.1, 11.2, 11.3, and 11.4, where *P* is the percentage of tested versus eligible students, by gender, university, disciplinary field, and high school type.

¹Please visit http://cae.org/images/uploads/pdf/CLA_Practice_Assessment.pdf for a sample assessment.

Table 11.1 TECO 2013 participation by gender

Gender	Eligible students	Tested students	<i>P</i>
<i>F</i>	13,468	3473	25.79
<i>M</i>	8404	2380	28.32
<i>F + M</i>	21,872	5853	26.76

Table 11.2 TECO 2013 participation by university

University	Eligible students	Tested students	<i>P</i>	Geographic area
MI University of Milan	2574	798	31.00	North
PD University of Padua	1918	549	28.62	North
PO University of Eastern Piedmont	506	319	63.04	North
UD University of Udine	448	287	64.06	North
BO University of Bologna	2645	368	13.91	Center
FI University of Florence	2457	691	28.12	Center
RM1 University of Rome “La Sapienza”	5808	1657	28.53	Center
RM2 University of Rome “Tor Vergata”	1080	183	16.94	Center
CA University of Cagliari	547	129	23.58	South
LE University of Salento	555	157	28.29	South
ME University of Messina	358	131	36.59	South
NA1 University of Naples “Federico II”	2976	584	19.62	South
ITA12 All 12 universities	21,872	5853	26.76	

11.3.3 Participants (2015)

The main differences between participants from TECO 2013 and TECO 2015 were the following:

- Students had to be in their third consecutive year enrolled at university.
- Students enrolled in a three-year first-cycle course must have acquired 75% of the basic and characterizing study credits required by the course class.
- Students enrolled in a single-cycle master’s course must have acquired at least 90 (from a total of 120) basic and characterizing study credits.

Tables 11.5, 11.6, 11.7, and 11.8 report the demographic summary of the TECO 2015 participating students compared to all eligible students, where *P* is the percentage of tested versus eligible students, by gender, university, disciplinary field, and high school type. Please note that the disciplinary field classification for the 2015 cohort is different than the 2013 classifications.

Table 11.3 TECO 2013 participation by disciplinary field

Disciplinary field	Eligible students	Tested students	<i>P</i>
Food and agriculture	361	139	38.50
Architecture	1136	272	23.94
Fine arts	412	64	15.53
Biology	805	256	31.80
Chemistry	278	106	38.13
Communication	510	131	25.69
Cultural heritage	523	142	27.15
Defense	0	0	0.00
Economics	1350	465	34.44
Pharmacy	1108	394	35.56
Philosophy	422	108	25.59
Education	431	128	29.70
Geography	252	53	21.03
Law	4319	874	20.24
Engineering	1219	463	37.98
Arts	681	190	27.90
Languages	1173	231	19.69
Mathematics, physics, and statistics	829	388	46.80
Medicine	2100	393	18.71
Dentistry	244	44	18.03
Political science	678	201	29.65
Psychology	1108	191	17.24
Sociology	368	79	21.47
History	235	57	24.26
Territory	1045	391	37.42
Veterinary science	285	93	32.63
Total	21,872	5853	26.76

Table 11.4 TECO 2013 participation by high school type

School type	Eligible students			Tested students			<i>P</i>		
	<i>F</i>	<i>M</i>	<i>F + M</i>	<i>F</i>	<i>M</i>	<i>F + M</i>	<i>F</i>	<i>M</i>	<i>F + M</i>
Not available	761	491	1252	162	141	303	21.29	28.72	24.20
Other school type	2387	692	3079	599	177	776	25.09	25.58	25.20
High school (Liceo)	8848	5621	14,469	2301	1594	3895	26.01	28.36	26.92
Professional institute	240	143	383	53	46	99	22.08	32.17	25.85
Technical institute	1232	1457	2689	358	422	780	29.06	28.96	29.01
Total	13,468	8404	21,872	3473	2380	5853	25.79	28.32	26.76

Table 11.5 TECO 2015 participation by gender

Gender	Eligible students	Tested students	<i>P</i>
<i>F</i>	17,011	3667	21.56
<i>M</i>	12,569	2655	21.12
<i>F + M</i>	29,580	6322	21.37

Table 11.6 TECO 2015 participation by university

University	Eligible students	Tested students	<i>P</i>	Geographic area	
BG	University of Bergamo	1245	239	19.20	North
MO	University of Modena and Reggio Emilia	1714	201	11.73	North
PD	University of Padua	5207	515	9.89	North
PO	University of Eastern Piedmont	638	187	29.31	North
PR	University of Parma	1300	562	43.23	North
TO	Polytechnic University of Turin	2232	287	12.86	North
UD	University of Udine	1397	273	19.54	North
VC	University of Insubria	778	216	27.76	North
PG	University of Perugia	1612	267	16.56	Center
RM2	University of Rome “Tor Vergata”	1123	492	43.81	Center
RmEU	University of Rome “Europea”	108	60	55.56	Center
RmL	University of Rome “LUISS”	1164	376	32.30	Center
SI	University of Siena	1138	411	36.12	Center
Sist	University for Foreigners of Siena	143	50	34.97	Center
UTIU	The International Telematic University Uninettuno of Rome	22	1	4.55	Center
UTMA	Telematic University G. Marconi of Rome	154	28	18.18	Center
BA	University of Bari	2643	492	18.62	South
CS	University of Calabria	1786	481	26.93	South
FG	University of Foggia	336	112	33.33	South
ME	University of Messina	1087	312	28.70	South
NA2	University of Naples “L’Orientale”	722	146	20.22	South
RCst	University for Foreigners of Reggio Calabria	72	50	69.44	South
SA	University of Salerno	1840	220	11.96	South
ITA23	All 23 universities	29,581	6323	21.38	

Table 11.7 TECO 2015 participation by disciplinary field

Disciplinary field	Eligible students	Tested students	<i>P</i>
Agriculture, forestry, and fishery	732	207	28.28
Architecture and building	1427	264	18.50
Arts	491	78	15.89
Business and administration	3520	875	24.86
Computing	475	126	26.53
Education science	2462	303	12.31
Engineering and engineering trades	3690	727	19.70
Environmental protection	221	77	34.84
Health	1267	278	21.94
Humanities	3821	785	20.54
Journalism and information	869	129	14.84
Law	2171	510	23.49
Life sciences	1092	348	31.87
Manufacturing and processing	140	2	1.43
Mathematics and statistics	367	119	32.43
Personal services	951	141	14.83
Physical sciences	555	159	28.65
Social and behavioral science	3567	676	18.95
Social services	480	106	22.08
Transport services	70	13	18.57
Veterinary	92	35	38.04
Total	28,460	5958	20.93

In addition to the reported demographic questions (Tables 11.1, 11.2, 11.3, 11.4, 11.5, 11.6, 11.7, and 11.8), all participating students were required to answer additional survey questions, including composition of the household, family socioeconomic status, off-site or working status, any form of financial support for studying, diploma and university grades, national or local admission test scores, their perception of whether they had acquired competencies in their course of study, and attendance regularity. Also, all participating students had to sign a waiver to allow ANVUR to use their data for research purposes, as required by the privacy guarantor.

11.3.4 Project Timeline

CAE and ANVUR assembled a project timeline (Table 11.9) for both TECO 2013 and 2015. The timeline illustrates the broad steps required for the entire collaboration between all participating organizations and was approximately parallel for both projects.

Table 11.8 TECO 2015 participation by high school type

School type	Eligible students			Tested students			P		
	F	M	F + M	F	M	F + M	F	M	F + M
Not available	741	642	1383	130	89	219	17.54	13.86	15.84
Other school type	3686	1059	4745	664	185	849	18.01	17.47	17.89
High school (Liceo)	9099	6820	15,919	2107	1567	3674	23.16	22.98	23.08
Professional institute	460	311	771	73	49	122	15.87	15.76	15.82
Technical institute	2461	3181	5642	481	613	1094	19.54	19.27	19.39
Total	16,447	12,013	28,460	3455	2503	5958	21.01	20.84	20.93

Table 11.9 TECO 2013 and 2015 timeline

Month	Tasks	Organization
March	PT and SRQ selection	ANVUR
April	Translation and adaptation	cApStAn/ANVUR/CAE
	Pretesting and cognitive labs	ANVUR
May	Begin administration	ANVUR/CINECA
June	Scorer training	CAE/ANVUR
July	End administration	ANVUR/CINECA
August	Score student responses	ANVUR
September	Data and item analyses	CAE/ANVUR
December/January	Deliver final report	CAE
January (2016 only)	Back translate student responses	cApStAn
	Score Italian student responses	CAE
March (2016 only)	Deliver updated report	CAE

11.3.5 CLA+ Task Selection

The CLA+ comprises a PT and a set of SRQs. In order to select the most appropriate PT and set of SRQs for the Italian version of CLA+ (TECO), ANVUR appointed a committee of guarantors (CG) to oversee the selection process, translation and adaptation, validation of the process, and psychometrics. Members of the CG were selected because of their expertise and experience in neuroscience, psychometrics, or assessment. There were four members of the CG for TECO 2013 and five for TECO 2015.

The CG examined all CAE's available PTs and SRQs and discussed potential modifications and adaptations to the test and scoring methodology with CAE. Some of the tasks were too US-centric (e.g., local politics, sports teams, curriculum, etc.), thus inappropriate for use in an international study. However, after a review of the available tests, a PT and a set of SRQs were selected by the 2013 CG for TECO 2013, and a different PT and set of SRQs were selected by the 2015 CG for TECO 2015. The choices made by the CGs were confirmed by CAE measurement scientists as the most appropriate tasks, given the international perspective of the study.

In TECO 2013, ANVUR decided to limit the SRQ section to 20 instead of 25 questions due to concerns with timing and content. The committee felt that the students, unfamiliar with the testing format, would find it difficult to complete 25 questions within the 30-min timeframe. Because the students in 2013 did not exhibit behavior indicating they found the testing format difficult, for TECO 2015, the full set of 25 SRQs were selected from CAE's bank of questions. As part of the experimental design for the 2015 study, a group of experts appointed by ANVUR and trained by CAE produced a set of 25 SRQs developed and written in Italian (SRQ ITA). The design of the 2015 study was to randomly assign each student either the Italian or the translated and adapted American subsection of the SRQs, yielding eight combinations (Table 11.10). In addition to receiving one of the eight SRQ forms, the students were also administered the translated and adapted PT.

Table 11.10 TECO 2015 SRQ form distribution

	SRQ Form							
	1	2	3	4	5	6	7	8
SQR	USA	USA	USA	USA	ITA	ITA	ITA	ITA
CRE	USA	USA	ITA	ITA	USA	USA	ITA	ITA
CA	USA	ITA	USA	ITA	USA	ITA	USA	ITA

11.3.6 Translation and Adaptation

Translation and adaptation of materials for international and cross-cultural assessment are very challenging endeavors (Geisinger 1994; Hambleton 2004; Wolf et al. 2015; Zlatkin-Troitschanskaia et al., Chap. 12 in this volume). The goal of the translation and adaptation process for these two studies was to localize the assessment to be consistent with the culture, history, and context of students' home country. This process ensured that TECO was analogous and equivalent to CLA+. CAE followed a similar process with nine participating countries in the OECD's AHELO feasibility study (Benjamin et al. 2012; Klein et al. 2013; OECD 2012b, c, 2013a).

Adaptations to the PT and SRQs to make the assessment more culturally appropriate were recommended by the CG, and CAE made the necessary changes. For example, the names of the cities used in the PT were adapted from American-centric names to more appropriate Italian city names. Other adaptations to the assessment, which were deemed to be too radical of a change, thus potentially affecting the validity of the equivalence between the two tests, were avoided.

Following the initial review by the CG for adaptations to the assessment, the translation of the PT and SRQs was conducted by independent organizations. In 2013, both INVALSI, with extensive experience in assessment at the secondary education level, and cApStAn, a professional translation and adaptation organization experienced with international assessment projects such as the PISA, independently performed the translations and translation verification and recommended a few additional adaptations. In 2015, cApStAn alone provided the translation and translation verification service. All the translations were carried out under the supervision of ANVUR and with the approval of CAE. In addition to the assessment itself, test operation documents such as test administration and scorer training were also translated.

11.3.7 Pretesting and Cognitive Labs

As part of best practice for all new assessments, ANVUR pretested the translated and adapted test and conducted a series of cognitive labs (Zucker et al. 2004). The main objectives of the pretesting and cognitive labs were to verify fidelity of the translation/adaptation to original constructs, confirm that the questions were

interpreted by Italian students with the original English meaning, and ensure that the translations and adaptations were not more difficult to read or understand than if they had originally been written in Italian.

For TECO 2013, the pretest was administered to 44 students at the University of Camerino. For TECO 2015, the pretest was given to 345 students from the University of l'Aquila. The pretesting process was intended to identify issues with administration of the assessment, detect any flaws or inconsistencies in the content, and evaluate the appropriateness of the translation and adaptation of the instrument. Additionally, pretesting allowed ANVUR to identify any possible residual problems and gaps in the demographic and contextual survey.

In addition to pretesting the larger cohort of students, CAE recommended that ANVUR conducts cognitive labs with a smaller sample of students. This was an attempt to collect qualitative data on the thought processes students engaged in while answering the questions on the assessment. The students were asked to "think aloud" while answering the questions on the assessment. They were allowed to explain their thought processes without interruption or corrective interventions from the interviewer. Additional questions intended to collect information about the process itself were posed at the end of the cognitive lab session.

The pretesting and cognitive labs showed that all students were able to read the translated and adapted text without difficulty. Additionally, one student identified some error and recommended lexical improvements for clearer understanding of the text. The errors were corrected, and the minor translation changes were implemented, yielding a final, validated, translated, and adapted assessment.

11.3.8 Administration

TECO was administered in the participating universities between May and July. The testing administration window was selected to ensure that the students took the exam after classes ended and before final exams were given. Because the students were all assessed during this time period, ANVUR was able to validate proper sampling and data collection protocols.

All TECO sessions were administered online in a proctored environment. One of the unique features of the PT is that cheating is challenging. For one, there are between six and eight documents for students to read, analyze, and synthesize. Moreover, students are required to construct a lengthy written response, for which there is not a single correct answer. For the SRQ section, the questions were randomly distributed to the students within each subsection, so students seated in close proximity were not given the same sequence of items. No testing irregularities were reported in either year, so ANVUR and CAE did not conduct any analyses (Wollack 1997) to detect cheating.

TECO was administered by a third-party collaborator, CINECA. CINECA is a consortium of Italian universities, research centers, and the Ministry of Universities and Research (MIUR). CINECA supports the research community by handling

projects such as large-scale test administrations and other computing and information system needs. Students' written responses to the PT, answers to the SRQs, and responses to the demographic survey were collected by CINECA. Once testing, scoring, and anonymization were completed by ANVUR, CINECA sent the dataset to CAE for data cleanup, analysis, and reporting. As part of the data cleanup, CAE calculated the average scores for each PT response and equated and scaled the Italian scores to the American version of the assessment.

Due to Italian privacy regulations, the universities were not authorized to receive identified (i.e., non-anonymized) results for individual students. This precluded universities from rewarding or incentivizing high-performing students. As a result of these restraints, ANVUR decided to provide students who participated in the studies, upon request, with individualized score reports. Moreover, universities could request the anonymized data not only of their own students but also for the entire sample of students from TECO 2013 and/or 2015.

11.3.9 Scorer Training and Scoring

Since the SRQs are scored objectively and dichotomously, all student results from this section of the assessment were machine-scored. CAE gave CINECA the answer key, and students' scores were computed by CINECA's system using CLA+ scaling equations. The PTs, however, needed to be hand scored using a six-point rubric across three sub-scales. The scoring of the PTs required an in-depth scorer training and verification process, which CAE led to ensure equivalency in scorer training methodology and calibration of the scorers.

For both TECO 2013 and 2015, ANVUR asked the rectors of the participating universities to appoint a lead scorer (LS). The LS at each university was responsible for recruiting scorers (university professors) within their institution, coordinating with the other LSs and ANVUR, and overseeing the final scoring of the student responses when discrepancies or issues were identified. ANVUR required LSs to be professors with academic authority and institutional influence, so that they could effectively implement the requirements of the study within their university. ANVUR selected an appropriate LS from each participating university, and the LS recruited a proportional number of scorers from within his or her institution.

Scorer training for the PT was extensive. For many of the participating scorers (professors), it was their first experience with scoring students' written responses using a standardized rubric and calibrating their scores with CAE's scoring experts and each other. The goal of the scorer training was not only to ensure consistent training but also to educate the scorers on the process. While competence and intellectual honesty are necessary to fairly assess the student, those characteristics in a scorer are insufficient to yield an unbiased score. In order to validate that the scoring process is as objective as possible, scorer training and scoring calibration are essential. Following the training and calibration exercise, as an additional check, student responses were double-scored and checked for scorer reliability.

The student responses were scored using a rubric that ranged from 1 to 6 for three sub-scores, APS, WE, and WM.² A score of N/A was assigned to students who did not answer the prompt or whose responses were off topic.

Scorer training of the scorers first began with an online training session with the ANVUR team and a select group of LSs. Measurement scientists from CAE led a half-day online scorer training meeting to orient the core scoring team to the scoring process. Following the online training, the core scoring team was given a homework assignment to score a set of 25 previously scored and verified student responses and submit their scores to CAE. Any scorers who needed additional training based upon the homework results had one-on-one skype sessions with CAE to calibrate their scores.

In June of each administration, ANVUR hosted a two-day, in-person scorer training for the core scoring team and all LSs from the participating universities. The initial online training and subsequent in-person trainings were conducted predominantly in English and used student responses that were written in English. Following the in-person training, the LSs then trained the scorers from their universities. The LS-lead trainings were conducted in Italian using Italian student responses and translated scoring materials (e.g., scoring rubric, scoring handbook). An additional Italian scoring guide, which provided detailed scoring response features and other instructions for each of the subscores (APS, WE, and WM), was developed by members of the LS team.

Once scoring commenced, the student responses were randomly and anonymously assigned to the scorers. Scorers completed their scoring tasks online using CINECA's scoring platform, which also allowed LSs to monitor the progress of their scorers.

In order to check for scorer reliability, for TECO 2013, 20% of the student responses were randomly selected and double-scored. Each student response that was selected was scored by two individuals. For any double-scored response that was inconsistent (i.e., the difference between the two total scores was greater than 3 points or the difference between two sets of subscores was greater than 2 points), INVALSI identified the pairs of scorers and checked to see if any scorers were consistently uncalibrated with the rest of the scorers. In total, only three out of 110 scorers were identified as uncalibrated and inconsistent, meaning scorer training was successful. The student responses that were scored by these three individuals were subsequently checked by INVALSI, who flagged any additional student responses that needed rescoring. All identified student responses were rescored, and the data file was revised by INVALSI before sending it to CAE. CAE used the average scores for the responses that were double-scored for all analyses.

For TECO 2015, all student responses were double-scored. Using the same scoring rules from 2013 for identifying inconsistent scorers, each flagged student response was scored by a third individual who was part of the ANVUR core scoring team. CAE computed the average scores for the closest two sets of scores for data

²Please see http://cae.org/images/uploads/pdf/CLA_Plus_Scoring_Rubric.pdf for the CLA+ scoring rubric.

analyses. The inter-rater correlation coefficients for each of the subscores ranged from $r = 0.80$ to 0.86 .

Lastly, CAE calculated the PT, SRQ, and TECO total scale scores for each student using the linear transformation equations used in the American version of the assessment.

11.4 Results

11.4.1 CLA+/TECO Test Level

TECO 2013 was completed by 5853 participants across 12 institutions. TECO 2015 was administered to 6245 students across 23 institutions. All students scoring N/A on the PT (meaning they did not respond to the prompt or their response was off-topic) were removed from the analyses. Table 11.11 contains the descriptive statistics for the two administrations by subsection. Please note that there were only 20 SRQs in TECO 2013 due to concerns with timing and content. Additionally, a set of SRQs developed entirely in Italian was administered in 2015. The scores were normally distributed for both sets of SRQs and the PT; however, the Italian items were more difficult than the American items.

For TECO 2015, because two sets of SRQ tests were administered, there were eight forms of the SRQs, using a combination of US and ITA items (Table 11.11). A one-way ANOVA to compare the mean difficulty level across the eight forms showed a significant difference in difficulty ($F(7, 6527) = 76.75; p < 0.0001$). The ITA items ($M = 9.69; SD = 3.34$) were more difficult than the US items ($M = 11.56; SD = 4.19$) (Fig. 11.1). As a result of the difference in mean scores across the eight forms, the SRQs were linearly equated to the US set and scaled. The subsequent analyses were conducted on these scale scores so that the data set could be analyzed as a single set of results.

Table 11.11 Descriptive statistics for the SRQs and PT

	TECO 2013		TECO 2015		
	SRQ	PT	SRQ_USA	SRQ_ITA	PT
Number of items	20	1	25	25	1
Number of students	5853	5853	938	870	6245
Min	0	3	0	0	3
Max	19	18	23	19	18
Mean	12.31	9.17	11.94	9.74	9.76
Median	13.00	9	12	10	9.5
St. Dev.	2.85	2.95	4.19	3.34	2.57

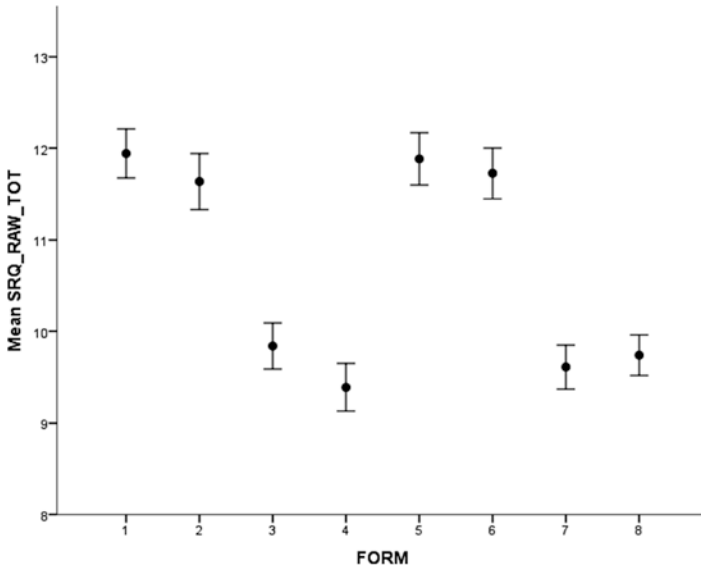


Fig. 11.1 Mean SRQ score with 95% CI error bars by test form

Table 11.12 Correlation coefficients of PT and SRQ scores

		PT_APS	PT_WE	PT_WM	PT_TOT	SRQ_TOT
2013	PT_APS	1.00				
	PT_WE	0.81**	1.00			
	PT_WM	0.61**	0.70**	1.00		
	PT_TOT	0.90**	0.93**	0.86**	1.00	
	SRQ_TOT	0.25**	0.25**	0.23**	0.27**	1.00
2015	PT_APS	1.00				
	PT_WE	0.85**	1.00			
	PT_WM	0.66**	0.74**	1.00		
	PT_TOT	0.92**	0.95**	0.87**	1.00	
	SRQ_TOT	0.28**	0.27**	0.23**	0.29**	1.00

**Correlation is significant at the 0.01 level (2-tailed)

Table 11.12 contains the correlation coefficients for the PT subscores (analysis and problem solving, writing effectiveness, and writing mechanics) and total PT score and SRQ score. The correlations between the SRQ and the PT subscores and total score, although statistically significant ($p < .01$), are unusually low ($r = .29$). The CLA+, administered domestically in the USA, typically has correlation coefficients of at least $r = .50$. The correlations across the PT subscores with each other are as expected. Since the PT and SRQs require different cognitive processes, this could be a possible explanation for the low correlation between the two sections.

11.4.2 University Level

Both TECO 2013 and 2015 were research studies that depended on voluntary participation from both universities and students within the universities. Thus, it is difficult when analyzing the data to adequately identify and correct for self-selection bias. Indeed, analysis conducted on TECO 2013 results (ANVUR 2014) showed that there was, in fact, a positive self-selection bias, which complicates the comparative assessment of disciplinary fields or universities characterized by very different participation indices (P) and TECO results. For example, it is not possible to assert that—under the condition of all other things being equal as regards all contextual variables—the University of Bologna (BO) had greater mean success in the TECO 2013 than the University of Eastern Piedmont (PO), due to very large differences in participation rates: under 14% for Bologna versus more than 63% for Eastern Piedmont.

Tables 11.13, 11.14, 11.15, and 11.16 report the mean and the participation indices (P) by university and disciplinary field for TECO 2013 and 2015. In addition to differences in the participation index, the mean scores also have a wide range.

Table 11.13 TECO 2013 participation index and mean scores by university

University	Geographic area	P	Mean PT	Mean SRQ	Mean TECO
MI	North	31.00	1036.55	1032.39	1034.53
PD	North	28.62	1024.14	1026.71	1025.49
PO	North	63.04	1017.21	972.55	994.92
UD	North	64.06	1020.45	1028.01	1024.28
BO	Center	13.91	1039.69	1055.61	1047.73
FI	Center	28.12	1016.02	1033.26	1024.71
RM1	Center	28.53	974.09	977.86	976.04
RM2	Center	16.94	984.34	978.00	981.22
CA	South	23.58	990.66	981.41	986.10
LE	South	28.29	940.89	938.68	939.92
ME	South	36.59	927.40	924.01	925.78
NA	South	19.62	971.17	959.57	965.43
ITA12		26.76	999.46	999.48	999.53

Table 11.14 TECO 2015 participation index and mean scores by university

University ^a	Geographic area	P	Mean PT	Mean SRQ	Mean TECO
BG	North	19.20	1002.92	1074.35	1038.69
MO	North	11.73	971.61	1087.06	1029.34
PD	North	9.89	983.17	1085.70	1034.43
PO	North	29.31	965.72	1034.17	1000.01
PR	North	43.23	956.45	1026.73	991.60

(continued)

Table 11.14 (continued)

University ^a	Geographic area	<i>P</i>	Mean PT	Mean SRQ	Mean TECO
TO	North	12.86	1008.72	1118.88	1063.83
UD	North	19.54	987.35	1060.28	1023.90
VC	North	27.76	981.13	1052.33	1016.71
AQ	Center	30.80	932.16	927.53	929.88
PG	Center	16.56	961.05	1026.37	993.74
RM2	Center	43.81	987.78	1056.58	1022.24
RmEU	Center	55.56	861.12	949.45	905.35
RmL	Center	32.30	1015.73	1091.79	1053.75
SI	Center	36.12	934.41	1023.50	979.00
Sist	Center	34.97	989.10	959.70	974.40
UTMA	Center	18.18	933.89	995.86	965.00
BA	South	18.62	943.87	989.31	966.58
CS	South	26.93	904.21	968.15	936.16
FG	South	33.33	900.38	968.38	934.38
ME	South	28.70	896.88	968.57	932.75
NA	South	20.22	978.24	993.40	985.78
RCst	South	69.44	747.18	861.65	804.51
SA	South	11.96	944.46	1036.82	990.68
ITA23		21.38	957.91	1028.18	993.06

^aFor privacy purposes, results of the single Uninettuno student are not shown

Table 11.15 TECO 2013 participation index and mean scores by disciplinary field

Disciplinary field	<i>P</i>	Mean PT	Mean SRQ	Mean TECO
Food and agriculture	38.50	981.22	986.72	984.01
Architecture	23.94	989.99	1021.73	1005.94
Fine arts	15.53	975.30	954.58	965.03
Biology	31.80	997.59	1015.21	1006.43
Chemistry	38.13	967.51	1023.29	995.45
Communication	25.69	989.09	966.36	977.81
Cultural heritage	27.15	986.12	969.73	977.99
Defense	0.00			
Economics	34.44	982.34	999.80	991.15
Pharmacy	35.56	979.59	971.19	975.45
Philosophy	25.59	1032.05	1004.23	1018.20
Education	29.70	932.83	873.38	903.28
Geography	21.03	985.51	882.23	933.96
Law	20.24	1021.76	997.59	1009.73
Engineering	37.98	980.10	1022.16	1001.20
Arts	27.90	1039.17	985.73	1012.51
Languages	19.69	1000.38	970.34	985.38
Mathematics, physics, and statistics	46.80	1027.57	1055.19	1041.43

(continued)

Table 11.15 (continued)

Disciplinary field	<i>P</i>	Mean PT	Mean SRQ	Mean TECO
Medicine	18.71	1057.48	1086.91	1072.25
Dentistry	18.03	1006.57	1023.75	1015.30
Political science	29.65	1015.05	997.13	1006.18
Psychology	17.24	1020.74	1038.62	1029.75
Sociology	21.47	986.28	929.61	958.04
History	24.26	1036.86	985.00	1011.02
Territory	37.42	938.69	932.54	935.70
Veterinary science	32.63	982.90	1025.14	1004.11
Total	26.76	999.46	999.48	999.53

Table 11.16 TECO 2015 participation index and mean scores by disciplinary field

Disciplinary field ^a	<i>P</i>	Mean PT	Mean SRQ	Mean TECO
Agriculture, forestry, and fishery	28.28	933.15	1014.10	973.67
Architecture and building	18.50	954.12	1048.33	1001.21
Arts	15.89	939.49	989.01	964.27
Business and administration	24.86	960.03	1029.84	994.94
Computing	26.53	937.94	1036.25	987.11
Education science	12.31	890.76	924.38	907.61
Engineering and engineering trades	19.70	976.43	1094.68	1035.59
Environmental protection	34.84	944.48	1055.29	1000.32
Health	21.94	984.46	1070.63	1027.58
Humanities	20.54	996.90	1026.94	1011.92
Journalism and information	14.84	943.95	1022.16	983.03
Law	23.49	969.24	1032.62	1000.97
Life sciences	31.87	932.14	1036.92	984.53
Mathematics and statistics	32.43	993.99	1077.83	1035.89
Personal services	14.83	910.99	965.76	938.35
Physical sciences	28.65	951.72	1072.94	1012.33
Social and behavioral science	18.95	968.19	1036.05	1002.12
Social services	22.08	843.47	914.67	879.12
Transport services	18.57	896.69	1041.00	968.85
Veterinary science	38.04	984.34	1040.54	1012.63
Total	20.93	959.40	1033.99	996.71

^aFor privacy purposes, results of the two students of the manufacturing and processing (MAN) disciplinary field are not shown

In the breakdown by gender (Table 11.17), female students show, on average, a lower participation index and lower scores, especially on the Scientific and Quantitative Reasoning section of the SRQs. The gender gap (to the disadvantage of female students) is particularly marked in the South, for both participation and test results (data not shown).

Table 11.17 TECO 2013 and 2015 participation index and mean scores by gender

Gender	TECO 2013				TECO 2015			
	<i>P</i>	Mean PT	Mean SRQ	Mean TECO	<i>P</i>	Mean PT	Mean SRQ	Mean TECO
<i>F</i>	25.79	1000.23	988.92	994.65	21.56	952.75	1006.39	979.6
<i>M</i>	28.32	998.33	1014.88	1006.66	21.12	965.05	1058.32	1011.69
<i>F + M</i>	26.76	999.46	999.48	999.53	21.37	957.91	1028.18	993.06

With respect to high school type, students who attend a “classical or scientific studies” high school (called “Liceo” in Italy) show a better performance on TECO compared to those coming from other types of institutions (Tables 11.18 and 11.19).

Moreover, students who have parents with “high cultural status,” an index that is used in Italy, have higher TECO scores. Students with at least one parent with a university degree or high school diploma, regardless of the father’s cultural status position, have higher mean scores than the grand mean of all students who participated. As expected, this effect is increased for students having both parents with a high cultural status. In general, we know that family status is predictive of the type of secondary school diploma a student will earn, the diploma grade, the course of study chosen at university, and cumulative university grade—in addition to directly predicting the results of his or her TECO score (Tables 11.20 and 11.21).

Of particular interest is the examination of the connection, or lack thereof, between the level of generic competences acquired during university studies (as perceived by the tested graduating students) and the level of performance on TECO. It would thus seem legitimate to conclude that students’ perception that they have acquired the right competences (expressed by more than 80% of the tested students in both years) is indicative only of high *customer satisfaction* but of nothing else of objective character (Tables 11.22 and 11.23).

11.4.3 Cross-Country Comparisons

In a cross-country comparison of students’ critical-thinking and written-communication skills, results show that the Italian students’ performance on TECO is roughly comparable to the results attained by their American counterparts for the PT (Table 11.24).

For the SRQ section, results varied. In 2013, Italian students outperformed American students, whereas in 2015, the opposite was true (Table 11.25). This could have been due to the increase in the number of SRQs in 2015 compared to 2013. So, on average, the students had more time per question in 2013 than in 2015.

Table 11.18 TECO 2013 mean scores by high school type

School type	<i>F</i>			<i>M</i>			<i>F + M</i>		
	Mean PT	Mean SRQ	Mean TECO	Mean PT	Mean SRQ	Mean TECO	Mean PT	Mean SRQ	Mean TECO
Not available	950.33	903.06	926.77	923.53	944.71	934.18	937.86	922.44	930.22
Other school type	988.70	964.94	976.92	1015.26	1012.11	1013.73	994.76	975.70	985.32
High school (Liceo)	1011.67	1008.90	1010.35	1016.18	1030.05	1023.17	1013.51	1017.56	1015.60
Professional institute	913.70	847.77	880.89	961.46	956.28	958.96	935.89	898.19	917.16
Technical institute	981.40	960.36	970.97	952.81	988.57	970.75	965.93	975.62	970.85
Total	1000.23	988.92	994.65	998.33	1014.88	1006.66	999.46	999.48	999.53

Table 11.19 TECO 2015 mean scores by high school type

School type	<i>F</i>			<i>M</i>			<i>F + M</i>		
	Mean PT	Mean SRQ	Mean TECO	Mean PT	Mean SRQ	Mean TECO	Mean PT	Mean SRQ	Mean TECO
Not available	900.37	956.02	928.22	956.78	1057.97	1007.42	923.29	997.45	960.40
Other school type	936.90	993.53	965.24	937.12	1047.64	992.41	936.94	1005.32	971.16
High school (Liceo)	971.65	1031.01	1001.35	986.47	1084.71	1035.59	977.97	1053.91	1015.96
Professional institute	870.07	935.48	902.84	931.67	1028.02	979.73	894.81	972.65	933.72
Technical institute	931.93	979.59	955.80	926.46	1022.23	974.37	928.86	1003.48	966.21
Total	954.61	1011.81	983.24	966.00	1064.61	1015.31	959.40	1033.99	996.71

Table 11.20 TECO 2013 mean scores by parent’s school level

Student’s parents with at least a baccalaureate degree	Tested students	Mean PT	Mean SRQ	Mean TECO
Both	1078	1022.55	1029.67	1026.19
One	1346	999.29	1006.08	1002.73
None	3429	992.26	987.39	989.90
ITA12	5853	999.46	999.48	999.53

Table 11.21 TECO 2015 mean scores by parent’s school level

Student’s parents with at least a baccalaureate degree	Tested students	Mean PT	Mean SRQ	Mean TECO
Both	798	981.91	1070.79	1026.35
One	1153	966.33	1052.72	1009.57
None	4007	952.92	1021.27	987.11
ITA23	5958	959.40	1033.99	996.71

Table 11.22 TECO 2013 mean scores by students’ self-assessment of adequacy of the competences acquired at university

“Competences acquired at university are adequate to perform well on the TECO?”	Tested students	Mean PT	Mean SRQ	Mean TECO
Not available	36	965.47	975.67	970.50
No	1137	1005.46	1004.77	1005.18
Yes	4680	998.26	998.37	998.39
ITA12	5853	999.46	999.48	999.53

Table 11.23 TECO 2015 mean scores by students’ self-assessment of adequacy of the competences acquired at university

“Competences acquired at university are adequate to perform well on the TECO?”	Tested students	Mean PT	Mean SRQ	Mean TECO
Not available	0			
No	1036	963.33	1037.98	1000.67
Yes	4922	958.57	1033.15	995.88
ITA23	5958	959.40	1033.99	996.71

Table 11.24 Descriptive statistics for the PTs for Italian vs. American students

		N	Mean	SD	Percentiles		
					25th	50th	75th
2013	ITA	5853	1000	200	852	989	1124
	USA	4380	1067	203	937	1070	1159
2015	ITA	972	1108	161	978	1105	1232
	USA	516	1102	180	1008	1128	1208

Table 11.25 Descriptive statistics for the SRQs for Italian vs. American students

		<i>N</i>	Mean	SD	Percentiles		
					25th	50th	75th
2013	ITA	5853	1000	200	908	1048	1119
	USA	4380	796	152	681	803	884
2015	ITA	972	1052	178	931	1060	1177
	USA	509	1126	183	1000	1138	1261

11.4.4 Back Translation of Italian PT Responses

CAE selected 25 Italian student responses that had perfect Italian scorer agreement to be back translated by cApStAn from Italian into English. The translations and adaptations maintained the authenticity of the student responses. For example, if the student made a grammatical error in Italian, a similar error in English was made. The adaptation also included changing the cities back to their original names (Clinton and Greenville) rather than keeping Borgorosso and Borgoverde as the city names.

The 25 translated and adapted student responses were initially scored by two CLA+ scorers. The responses were mixed in with 25 American student responses. The scorers were blind to the fact that half of the student responses were back-translated Italian student responses. A third scorer was brought in to score five of the 25 responses because there was a difference of greater than 2 points between the initial two scorers. The two closest scores were averaged, and that score was used for subsequent analyses. The inter-rater reliability as measured by the Pearson correlation between the two total PT scores was $r = 0.97$; $p < 0.001$.

The correlation between the average total PT score for teams of American and Italian scorers was $r = 0.76$, $p < 0.01$. The Italian and American scorers had different mean scores for the 25 student responses (M.ITA = 9.72, SD.ITA = 5.13; M.USA = 11.06, SD.USA = 3.80). However, the average difference between the Italian scorers and the American scorers ($M = 1.34$; $SD = 3.33$) was not found to be significant ($t_{24} = 2.02$; $p = 0.055$; Fig. 11.2). This result provides evidence that the scoring process, which includes scorer training, is valid and provides comparable results between the Italian and American teams.

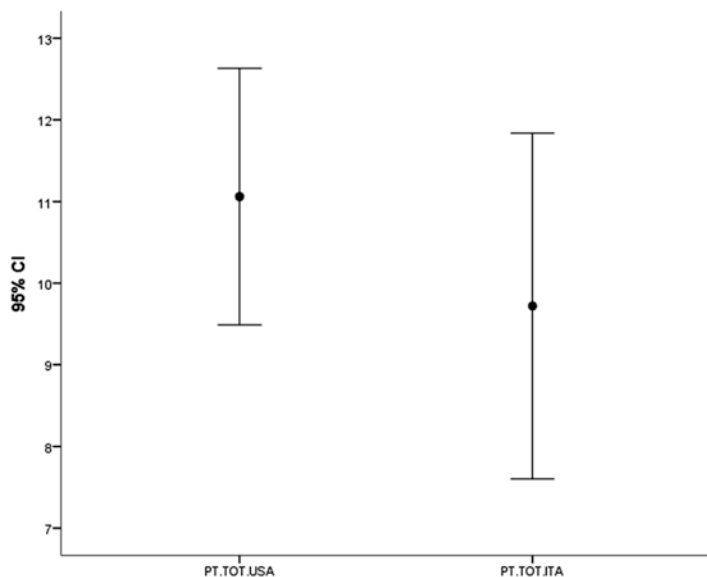


Fig. 11.2 Mean PT total score with 95% CI error bars by scoring team, $N = 25$ student responses, $t_{24} = 2.02$; $p = 0.055$

11.5 Discussion

When ANVUR undertook the feasibility study using CLA+, the purpose was to assess Italian students' generic skills and conduct cross-country benchmarking. The results show that the CLA+ can indeed be used to assess these skills and that the Italian students' performance on TECO is roughly comparable to the results attained by their American counterparts. There are, however, puzzling questions to resolve and a few methodological issues to further investigate. For example, the correlations between the two sections during both administrations of TECO were much lower than for the American students.

There are several hypotheses and one possible explanation regarding these relatively low correlation values. The first is that perhaps Italian students are not used to taking standardized tests, let alone PTs, so the correlation between the PT and SRQ scores is lower than expected. Table 11.24 shows that performance on the PTs is approximately equivalent for Italian and American students, but this is not the case for the SRQs (Table 11.25). However, the difference in performance is not huge. We know from recent results reported in the PIAAC study that there is variation in performance on cross-national standardized tests and that American students are not the top performers across all nations (OECD 2013b). We also know, based upon results of the AHELO feasibility study, that results on PTs vary from country to country (Zahner and Steedle 2014). However, we do not have benchmarking information of the CLA+ for other nations at this point. Italy's cross-national standing on the CLA+ will not be available until more countries have adopted the assessment.

A second hypothesis regarding the low correlation between the PT and SRQ scores is that the writing subscores from the PT might not be associated with the APS subscore on the PT and that the APS subscore on the PT might also not be associated with the APS skills measured by the SRQs. The correlations for TECO 2013 between APS and the writing subscores on the PT, as shown in Table 11.12, are not low ($r = 0.81$ [2013] and 0.85 [2015] for writing effectiveness and $r = 0.61$ [2013] and 0.66 [2015] for writing mechanics); however, the correlation between just the APS subscore and the SRQs as shown in Tables 11.12 and 11.26 is low.

Table 11.26 TECO 2013 correlation coefficients PT_APS and SRQ subscores ($N = 5853$)

	PT_APS	SRQ_CRE	SRQ_CA	SRQ_SQR	SRQ_TOT
PT_APS	1.00				
SRQ_CRE	0.20**	1.00			
SRQ_CA	0.16**	0.23**	1.00		
SRQ_SQR	0.15**	0.24**	0.21**	1.00	
SRQ_TOT	0.24**	0.77**	0.62**	0.68**	1.00

**Correlation is significant at the 0.01 level (two tailed)

One possible explanation of the low correlations and the equal average proportional means of the subsections is that each of the subsections requires different analysis and problem-solving skills and may be indicative of knowledge or instructional differences for the Italian students. Although no interaction was found between the topic of a PT and a student's college major for the American version of the CLA (Steedle and Bradley 2012), analyses of the interaction between specific subsections of the CLA+ SRQs and college majors have never been studied. This is due to the fact that the correlations between the three subsections of the CLA+ SRQs are high for the American students. For the Italian students though, this is something that might contribute to what is being observed. Perhaps engineering students perform better on the SQR section, students majoring in the humanities perform better on the CRE section, and students majoring in law and political science perform better on the CA section. Coursework and instruction for the Italian and American students differ in that there is less focus on general studies for the Italians and more on their specific content areas. It may be interesting to analyze whether there is an interaction between student majors and the SRQ subscores.

For TECO 2015, correlations for total and subscores of the PT and SRQs indicate that correlations were strong within PT subsections, but this was not the case within SRQ subsections (Table 11.27). This also suggests that the subsections of the SRQs are in fact measuring different aspects of analytic reasoning and evaluation skills.

Table 11.27 TECO 2015 Correlation coefficients—PT and SRQ and subsections for Italian students

	PT TOT	APS	WE	WM	SRQ TOT	SQR	CRE	CA
PT total	1.00							
APS	0.91**	1.00						
WE	0.94**	0.83**	1.00					
WM	0.87**	0.67**	0.75**	1.00				
SRQ total	0.29**	0.33**	0.31**	0.27**	1.00			
SQR	0.20**	0.23**	0.21**	0.18**	0.70**	1.00		
CRE	0.22**	0.27**	0.26**	0.22**	0.79**	0.23**	1.00	
CA	0.19**	0.17**	0.17**	0.14**	0.56**	0.11**	0.35**	1.00

**Correlation is significant at the 0.01 level (two tailed)

Interestingly, at the institutional level, the two sections are much more highly correlated (Table 11.28), indicating that performance on the TECO is as expected when the data are aggregated to the institutional level. Institutions with students who have high PTs scores also have students with high SRQ scores and vice versa. The exception to this observation is with the 2015 CA set. Even when aggregated, the correlations between CA items and the other two subsections of the SRQs are low. The items that were developed in Italian and administered in 2015 were not exactly aligned to the construct due to differences in the item development teams. The American items were developed using CAE-trained measurement scientists or item developers who were specifically trained to write items to the CLA+ construct. These individuals are not university professors. Whereas the Italian items were developed by a team of expert educators, who were not thoroughly familiar with the CLA+ construct. This difference could have contributed to the observed difference in the correlations between 2013 and 2015.

Table 11.28 TECO 2013 and 2015 correlation coefficients for PT and SRQ and subsections at the institutional level; *N* (2013) = 12 institutions; *N* (2015) = 23 institutions

		PT TOT	SRQ TOT	SQR	CRE	CA
2013	PT total	1.00				
	SRQ total	0.93	1.00			
	SQR	0.90**	0.94**	1.00		
	CRE	0.82**	0.96**	0.85**	1.00	
	CA	0.91**	0.90**	0.79**	0.79**	1.00
2015	PT total	1.00				
	SRQ total	0.68**	1.00			
	SQR	0.68**	0.97**	1.00		
	CRE	0.54**	0.97**	0.94**	1.00	
	CA	0.73**	0.45*	0.31*	0.27**	1.00

*Correlation is significant at the 0.05 level (two tailed)

**Correlation is significant at the 0.01 level (two tailed)

11.5.1 Solutions and Recommendations

Regardless of the results of the comparison between the two groups of students, there may be some merit to the hypothesis that the Italian students were not as familiar with standardized tests or PTs as American students. One recommendation would be to develop a practice assessment for the students, so they could familiarize themselves with PTs and standardized assessments as a whole. For AHELO, CAE developed a “mini-PT” (OECD 2011) and had it translated and adapted into seven languages as part of the feasibility study.

At the individual student level, results can be used to connect graduating students with potential employers. CAE has the CLA+ Career Connect, which awards digital badges based upon students’ performance on CLA+. These badges were a result of a standard-setting study that established five levels of mastery (Zahner 2014). Students can use these badges and send the results to employers who are seeking qualified individuals in the workforce. This has the advantage of increasing diversity in the workforce, since individuals from underrepresented groups (e.g., in the USA, it is certain races/ethnicities; in Italy, it could be based on disciplinary field or region) seeking employment can showcase their skills.

11.5.2 Future Research Directions

It is crucial for ANVUR to assess and certify the generic competences acquired by university students. Specifically, the research should investigate the predictive validity of TECO on students’ success after university. For example, what are the outcomes for the students in terms of employment, socioeconomic status, and analogical transfer (Gick and Holyoak 1980, 1983) of these important skills to novel situations? In the USA, these predictive validity studies have already been started, and results indicate that the CLA+ does predict positive post-university outcomes as measured by salary, employment, and enrollment in graduate school (Zahner and James 2016).

More generally, the methodology on international assessment could be improved upon through additional research. Currently, there are similar investigations occurring internationally. For example, there is a study of student learning gains in the UK where five institutions are longitudinally following students as they progress through university and measuring, among other variables, their critical-thinking and written-communication skills. Similarly, colleagues from Modeling and Measuring Competencies in Higher Education (KoKoHs) in Germany have conducted a feasibility study measuring students’ generic skills in Germany (see also Chap. 12, in this volume). Once the individual projects have been completed, the idea is to have data that can be used for cross-cultural comparisons between Italy, Germany, the UK, and the USA.

As a next step, international experts in tertiary education, assessment, and international, cross-country research could be convened to develop and refine instruments that are appropriate for multiple cultural contexts. This proposal is currently being implemented by CAE and the OECD. The two organizations are collaborating on a new initiative to launch the CLA+ International, Programme for Tertiary

Assessment, a large-scale international assessment of generic skills. This program will follow many of the same translation, adaptation, scorer training, and scoring protocols outlined in this chapter. Additionally, a standard-setting validation study will occur following the first administration. Representatives from the higher-education sector and industry from each participating country will validate the levels of mastery (Zahner 2014) and potentially establish badges for students seeking employment.

11.6 Conclusion

The collaboration between ANVUR and CAE was an important first step in formally assessing generic skills in higher education in an international context. Much of the methodology was followed based upon CAE's experience in the OECD's AHELO feasibility study. However, for the generic skills strand in AHELO, the assessment used was pieced together using two different assessments aligned to two different construct definitions of critical thinking (Klein et al. 2013). Thus, results were less than optimal. This study was the first of hopefully many studies occurring internationally measuring student learning gains and outcomes in tertiary education.

Overall, results from these studies indicate that it is feasible to translate and adapt a performance-based assessment to measure critical-thinking and written-communication skills of university students in Italy. It also is possible to conduct cross-country comparisons of these skills. The results indicated that Italian students' performance, for the most part, was comparable to their American counterparts. In 2013, the Italian students also significantly outperformed the American students on the SRQs, indicating that familiarity with this type of item is not an issue. This difference changed in 2015, potentially because 25 items were administered instead of 20.

International assessments are challenging (Blömeke et al. 2013; Chap. 10, in this volume), but these studies demonstrated that it is indeed feasible to measure critical-thinking and written-communication skills using a performance-based assessment. Evidence from these studies strengthens the case for assessing students in the tertiary systems internationally. The partnership established between the OECD and CAE is a testament to this.

References

- ANVUR. (2014). *Assessing the generic competencies acquired by students graduating from Italian universities*. Retrieved from Rome, Italy.
- Benjamin, R., Klein, S., Steedle, J. T., Zahner, D., Elliot, S., & Patterson, J. A. (2012). The case for generic skills and performance assessment in the United States and international settings. *CAE – Occasional Paper*. Retrieved from http://www.collegiatelearningassessment.org/files/The_Case_for_Generic_Skills_and_Performance_Assessment_in_the_United_States_and_International_Settings.pdf
- Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., & Fege, J. (2013). *Modeling and measuring competencies in higher education*. Rotterdam: Springer.

- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*(4), 304.
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*(3), 306–355.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1), 1–38.
- Hambleton, R. K. (Ed.). (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Psychology Press.
- Hart Research Associates. (2006). *How should colleges prepare students to succeed in today's global economy? – Based on surveys among employers and recent college graduates*. Retrieved from Washington, DC.
- Kahl, S. (2008). *The assessment of 21st century skills: Something old, something new, something borrowed*. Paper presented at the Council of Chief State School Officers 38th National conference on Student Assessment, Orlando, FL.
- Klein, S., Zahner, D., Benjamin, R., Bolus, R., & Steedle, J. T. (2013). *Observations on AHELO's generic skills strand methodology and findings*. New York: Council for Aid to Education.
- Levy, F., & Murnane, R. J. (2004). Education and the changing job market: An education centered on complex thinking and communicating is a graduate's passport to prosperity. *Educational Leadership, 62*(2), 80–83.
- OECD. (2011). *AHELO feasibility study progress report*. Retrieved from Paris, France.
- OECD. (2012a). *AHELO feasibility study report*. Retrieved from Paris, France.
- OECD. (2012b). *AHELO feasibility study report*. Retrieved from Paris, France.
- OECD. (2012c). *Assessment of higher education learning outcomes. Feasibility study report. Volume1 – Design and implementation*. Retrieved from <http://oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf>
- OECD. (2013a). *Assessment of higher education learning outcomes. Feasibility study report. Volume3 – Further insights*. Retrieved from <http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume3.pdf>
- OECD. (2013b). *OECD skills outlook 2013: First results from the Survey of Adult Skills*. Retrieved from Paris, France.
- OECD. (2016). *PISA FAQ*. Retrieved from <https://www.oecd.org/pisa/aboutpisa/pisafaq.htm>
- Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5–15.
- Steedle, J. T., & Bradley, M. (2012). *Majors matter: Differential performance on a test of general college outcomes*. Paper presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada.
- Wolf, R., & Zahner, D. (2015). Mitigation of test bias in international, cross-cultural assessments of higher-order thinking skills. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development*. Hershey: IGI Global.
- Wolf, R., Zahner, D., & Benjamin, R. (2015). Methodological challenges in international comparative post-secondary assessment programs: Lessons learned and the road ahead. *Studies in Higher Education, 1*–11. <https://doi.org/10.1080/03075079.2015.1004239>
- Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*, 307–320.
- Zahner, D. (2014). *CLA+ standard setting study final report*. Retrieved from New York, NY. http://cae.org/images/uploads/pdf/cla_ss.pdf
- Zahner, D., & James, J. K. (2016). *Predictive validity of a critical thinking assessment of post-college outcomes*. Paper presented at the American Educational Research Association, Washington, DC.
- Zahner, D., & Steedle, J. T. (2014). *Evaluating performance task scoring comparability in an international testing program*. Paper presented at the American Educational Research Association, Philadelphia, PA.
- Zucker, S., Sassman, C., & Case, B. J. (2004). *Cognitive labs*. San Antonio, TX: Harcourt Assessment. Retrieved, 10(11), 2006.

Chapter 12

Adapting and Validating the Collegiate Learning Assessment to Measure Generic Academic Skills of Students in Germany: Implications for International Assessment Studies in Higher Education



Olga Zlatkin-Troitschanskaia, Miriam Toepper, Dimitri Molerov, Ramona Buske, Sebastian Brückner, Hans Anand Pant, Sascha Hofmann, and Silvia Hansen-Schirra

Abstract Starting in 2015, a German research team from the program Modeling and Measuring Competencies in Higher Education (KoKoHs), in collaboration with the US Council for Aid to Education (CAE), adapted and validated the Collegiate Learning Assessment (CLA+) for the German language and cultural context to measure generic higher-order cognitive skills of university students and graduates in Germany. In this chapter, the conceptual and methodological background, the framework of the adaptation and validation study, as well as preliminary results are presented. Finally, findings are discussed critically, and future challenges and perspectives are explored.

12.1 Relevance and Background

Globalization, digitalization, and demographic change are current challenges in the societies, labor markets, and educational systems in most member countries of the Organization for Economic Co-operation and Development (OECD). Policy-driven

O. Zlatkin-Troitschanskaia (✉) · M. Toepper · R. Buske · S. Brückner · S. Hofmann
S. Hansen-Schirra
Johannes Gutenberg University, Mainz, Germany
e-mail: lstroitschanskaia@uni-mainz.de; miriam.toepper@uni-mainz.de;
buske@uni-mainz.de; brueckner@uni-mainz.de; s.hofmann@uni-mainz.de;
hansenss@uni-mainz.de

D. Molerov · H. A. Pant
Humboldt-Universität zu Berlin, Berlin, Germany
e-mail: molerov@hu-berlin.de; hansanand.pant@hu-berlin.de

reform strategies aimed at narrowing down the existing skill gaps between labor market demands and skill levels of students and graduates. The OECD skills strategy and the survey of adult skills in the OECD Program for the International Assessment of Adult Competencies (PIAAC) have gained international attention (OECD 2016). In higher education, prominent reform strategies such as the Bologna reform in Europe have raised questions regarding the individual and societal returns on higher education. There is a growing need for valid performance-based assessments of higher-order skills that can be used with different groups of students from different countries (see also Shavelson et al., Chap. 10 in this volume). One reason for this can be seen in the current internationalization and harmonization trends in higher education systems with regard to the bachelor-master study model, which have resulted in students becoming increasingly mobile between universities in different countries.

Student learning outcomes (SLOs) have been defined in national and international frameworks in order to manage the accreditation of degree courses and institutions in higher education (e.g., European Qualifications Framework (EQR), European Commission 2015, and the German Qualifications Framework (DQR)). At the institutional level, SLOs as the output of higher education have been defined in study program regulations and module descriptions. However, neither the certificates of academic achievement based on SLO specifications that have been established nationally or internationally nor various existing institutional ranking models have been based on suitable, psychometrically sound methods of assessment. On the contrary, grades and certificates are hardly comparable between higher education institutions even at the national or local level (Zlatkin-Troitschanskaia et al. 2017). Hence, national and international comparative assessment studies are becoming more relevant. These developments over the last decade have emphasized the importance of SLO assessments and the demand to measure SLOs in higher education in a valid, reliable, and fair manner (Zlatkin-Troitschanskaia et al. 2016b; see also Coates 2014, Coates, Chap. 1 in this volume).

Challenges specific to higher education such as high international and national diversity of degree courses, study programs, and institutions make developing and implementing SLO assessments in higher education and in particular assessments of students' generic higher-order cognitive skills a highly complex and multidimensional task. In most OECD countries, the importance of twenty-first century generic skills such as critical thinking, problem solving, quantitative and qualitative reasoning, analytical reasoning, information literacy, and digital literacy are recognized (cf. Alexander, Chap. 3 in this volume). Nonetheless, the increasing importance of such skills is undisputed in international educational practice and research. They are supposed to be a high priority for succeeding in knowledge-based economies, addressing judgments, decisions, and challenges in everyday life, and being an engaged citizen of a globalized world and are therefore necessary for individuals' lifelong learning (e.g., OECD 2014; Shavelson et al., Chap. 10 in this volume).

In order to provide a comprehensive overview of the research and developments in the field of competency assessment in higher education, the KoKoHs research team (Zlatkin-Troitschanskaia et al. 2016b) conducted a broad and detailed docu-

ment analysis, which included systematic literature and database searches and qualitative content analyses from 2010 to 2016. This review presented that grades across institutions are incomparable. The existing assessments are, for the most part, only suitable as higher education admission tests, for gathering data on individual learning opportunities and as subjective measures (Zlatkin-Troitschanskaia et al. 2015, 2017). Overall, the review's results suggested that the relevance of SLO assessments in higher education is continuously increasing, thanks to their potential to be used for multiple purposes and to provide multi-perspective, evidence-based information for diverse stakeholders (see, e.g., Spiel and Schober, Chap. 4 in this volume).

In Germany, the Federal Ministry of Education and Research (BMBF) has established a national research program on "Modeling and Measuring Competencies in Higher Education" (KoKoHs). The first funding phase (2011–2015) involved 24 collaborative projects comprising approximately 70 individual projects conducted by almost 220 researchers, focusing on modeling and measuring domain-specific and generic competencies in higher education.¹ In the next funding phase, which runs between 2016 and 2020, the new KoKoHs program focuses on "Validations and Methodological Innovations." The KoKoHs researchers build on the newly developed models, instruments, and findings, and validate assessments in greater depth according to the Standards of Educational and Psychological Testing ("the Standards," American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) 2014) and expand existing models and assessment instruments to be used in different study domains or for measurement over time. International connectivity and compatibility of assessments have been an important aim of the KoKoHs program as well (e.g., Brückner et al. 2014). Many KoKoHs project teams are eager to discover international best practice models and to adapt and validate more innovative international approaches for use in German higher education (e.g., the WiWiKom project, Zlatkin-Troitschanskaia et al. 2014, Brückner and Zlatkin-Troitschanskaia, Chap. 6 in this volume).

With focus on assessing generic skills, an increase in research efforts can also be observed at the international level (Zlatkin-Troitschanskaia et al. 2016b). The OECD's feasibility study Assessment of Higher Education Learning Outcomes (AHELO) was an initial approach to internationally assess SLOs in higher education (OECD 2013; Tremblay 2013). In addition to measuring domain-specific competencies in engineering and economics, AHELO employed the Collegiate Learning Assessment (CLA) to assess generic skills. Based on the experiences from AHELO, the US Council for Aid to Education (CAE) developed a new test, the CLA+, as a performance assessment which measures students' generic skills at the level of

¹ The outcomes of the KoKoHs research initiative, which also gave the basis for this study, included 40 competency models and more than 100 measuring instruments. The assessments were carried out with altogether more than 50,000 students at more than 220 higher education institutions throughout Germany to gather evidence of their psychometric quality (Zlatkin-Troitschanskaia et al. 2016a, b).

higher education in the United States (CAE 2013). So far, the CLA+ has been adapted for use in Italy and the United Kingdom. Currently, the CAE in cooperation with the OECD has launched a new program, CLA+ International, to further develop and expand the work on an international level (CAE 2015, Zahner and Ciolfi, Chap. 11 in this volume).

A meta-analysis by Zlatkin-Troitschanskaia et al. (2016b) showed that no German-language instruments for assessing performance exist that meet academic requirements for measuring university students' generic higher-order cognitive skills. Therefore, starting in 2015, a German research team from the KoKoHs program collaborated with the CAE to adapt and validate the CLA+ for the German language and cultural context to measure such skills of higher education students and graduates in Germany.

12.2 Aims and Framework of the German Adaptation and Validation Study

12.2.1 Goals

The goal of the German study was to enable the assessment of generic higher-order cognitive skills in Germany by adapting and validating the CLA+ for a German context while also aiming to ensure the international compatibility and comparability of the adapted assessments and results. In this sense, this study seeks to contribute substantially to the development of assessments of and research on university students' generic skills in Germany. The additional research challenge was to carry out the adaptation and validation in a way that the underlying concept and assessment framework of generic higher-order cognitive skills would be aligned with those established in other countries using the CLA+ (so far, the United States, Italy, and the United Kingdom). In all interpretations in the adaptation and validation process, the team aimed for functional equivalence between the German and the US versions (on functional equivalence, see Braun 2006).

To achieve these goals, the German study comprised four major milestones:

1. Translating the US test instrument into German and adapting it to the German culture to obtain a localized German instrument
2. Validating the German instrument comprehensively for use in higher education in Germany according to the Standards (AERA, APA, and NCME 2014)
3. Based on the validation results, exploring the need for further development and adaptation
4. In collaboration with the CAE team and possibly partners in Italy and the United Kingdom, conducting international comparability analyses

Adapting and validating an educational assessment is a complex and multifaceted task. In the German study, in order to ensure that the adapted instrument is of

high quality, the translation, adaptation, and validation processes were based on the Test Adaptation Guidelines (TAG) by the International Test Commission (ITC 2016; Coyne 2000; Hambleton 2001) and the Standards (AERA, APA, and NCME 2014). The TAG provide a rough orientation on appropriate framework conditions for adaptations and were specified for this project. The Standards provide general guidance on the validation of (1) test content, (2) response processes, (3) internal test structure, and (4) relations of the assessed construct to other variables (AERA, APA, and NCME 2014). To meet the validity criteria related to the (1) test content of the Standards, which correspond with the content criteria in the TAG, the German team had to ensure that the constructs of generic higher-order cognitive skills were conceptualized and understood in a similar way in Germany and the United States.

To this end, the theoretical concepts and models underlying the CLA+ tests by the CAE in order to validate it for Germany have been explored (Sect. 2.2). The test instruments were then translated and adapted (Sect. 3.1). The validation analyses so far have included curricular analyses, expert panels, and lecturers' online ratings (see Sect. 3.2). In addition, the (2) cognitive requirements and response processes were analyzed using cognitive interviews with students (Sect. 3.3).² Overall, in the German study, a systematic adaptation and validation framework were employed to determine whether the adapted assessment enables a valid measurement of generic higher-order cognitive skills among students and graduates in higher education in Germany. The next step will include preliminary comparability analyses with data from other countries (see milestone 4).

12.2.2 Study Framework

The term higher-order cognitive skills is not defined in a uniform way, and diverse conceptualizations and conceptual frameworks can be found in the research literature (e.g., an overview in Liu et al. 2014; Pellegrino and Hilton 2012). For example, Wheeler and Haertel (1993) conceptualized higher-order skills by determining two contexts in which these skills are employed: (a) situations where thought processes are needed for solving problems and making decisions in everyday life and (b) contexts where mental processes can be applied that have to be developed by formal instruction, including processes such as comparing, evaluating, and justifying. For both contexts, being able to employ higher-order skills is perceived as crucial in a knowledge-based society and digital world (see also Alexander, Chap. 3 in this volume). This kind of conceptualization is commonly accepted in international research, and the first context has served as a starting point for international assessment programs (Forster 2004). While the term higher-order skills refers to a very broad range of domains, the CLA+ aims to measure specific aspects (CAE 2013). The CLA+ assessments rubrics and the constructs have been developed to

²Further analyses of (3) the internal test structure and (4) relations to other variables will be conducted after the first administration of the test in the field.

holistically assess analytical reasoning and problem solving (Zahner and Ciolfi, Chap. 11 in this volume).

There are many approaches in measuring SLOs in higher education, such as self-report surveys of learning, multiple-choice tests, or short-answer tests (Zlatkin-Troitschanskaia et al. 2016b). However, the underlying concept of higher-order skills refers to real-life decision making and judgment, which should be reflected as closely as possible in the assessment format (Shavelson 2013; Shavelson et al., Chap. 10 in this volume). According to the literature on international studies on cognitive dispositions, such skills should be assessed mainly via complex item formats that present authentic cases with an adequate and meaningful action-oriented situational context from real life (e.g., Shavelson et al. 2015). Various studies recommend the use of different item formats for the assessment of different aspects of higher-order cognitive skills (e.g., Herl et al. 1996; Ruiz-Primo and Shavelson 1996; Snow 1993).

The CLA+ includes different case-based task formats and both complex performance tasks (PT) and selected-response questions (SRQs) administered on a computer. The PT consists of a short frame scenario and an additional document library where further information of varying relevance is presented. To respond, test takers are prompted to use the information and write a text (e.g., a report). The PT is designed to measure three dimensions: problem solving and analysis, writing effectiveness, and writing mechanics. The second task format, the SRQs, also present a situational context and prompt test takers to choose one correct answer from a selection of four to five options. The SRQs items are designed to assess three additional dimensions: scientific and quantitative reasoning, critical reading and evaluation, and the ability to criticize an argument. The length of the test is limited to 60 min for the PT and 30 min for the SRQs (see also Zahner and Ciolfi, Chap. 11 in this volume).

An overview of the project steps is provided in Table 12.1.

Table 12.1 Overview of the German study

Spring 2015	Selection of tasks (PT 1 and 25 SRQs) for the German study
Summer 2015	Workshop with CAE's developers of the CLA+, including scorer training
Summer 2015	Meeting with colleagues from Italian National Agency for the Evaluation of the University and Research Systems (ANVUR)
Summer/autumn 2015	Agreement on translation guidelines between CAE, German team and translation agency cApStAn
Autumn 2015	Translation by cApStAn (PT 1, 25 SRQs, test instructions, scoring guidelines)
Autumn 2015	Review and revisions of translation by German team and first adaptation round for PT1
Autumn 2015	Curricular analyses
Winter 2015/16	Expert workshop I: Group discussion with 10 national experts from different fields of studies
Winter 2015/16	Second adaptation round by German team for PT 1
Winter 2015/16	Expert workshop II: Group discussion with 10 national experts from different fields of studies

(continued)

Table 12.1 (continued)

Winter 2015/16	Third adaptation round by German team for PT1
Winter 2015/16	Translation of PT 2 by cApStAn
Winter 2015/16	Review and revisions of translation by German team and first adaptation round for PT2
Winter 2015/16	Expert workshop III: Group discussions with 3 translation experts
Spring 2016	Second adaptation round for PT 2 by German team
Spring 2016	Meeting with colleagues from UK's Learning Gain Program (representatives from the Centre for Excellence in Learning and Teaching (CELT))
Spring/summer 2016	Ten cognitive interviews with students (PT 1 and 2)
Spring/summer 2016	Localization of PT 2 by the German team
Autumn 2016	Back translation of localized PT 2 and review by CAE
Winter 2016	Online rating by 12 lectures (25 SRQs)
Winter2016/ spring 2017	Ten cognitive interviews with students with localized PT 2
Spring/summer 2017	Further analyses and documentation of results
Summer/autumn 2017	Comparison of the original version from the U.S. and the adapted test versions from the UK, Germany, and Italy
	Exchange of data and further cross-cultural comparative analyses

12.3 Project Overview and Preliminary Results of the Validation

12.3.1 Translation and Adaptation

In addition to the TAG, the Translation, Review, Adjudication, Pretesting, and Documentation (TRAPD) process then followed (Harkness 2003) – a standard process used when adapting international assessments and surveys. TRAPD is a process approach used to ensure that the test is reviewed, revised, and appraised by a variety of experts on its content, methodology, and translation (Harkness 2003, for a discussion of each step, see also the Cross-Cultural Survey Guidelines by Mohler et al. 2016; see also Behr and Shishido 2016).

The CLA+ was translated into German by cApStAn; a translation service provider specialized in the translation and adaptation of international educational and psychological tests.³ Linguistic supervision, translation reviewing, and quality assurance were provided by team members from the Faculty of Translation Studies, Linguistics, and Cultural Studies at Johannes Gutenberg University Mainz. The

³The company had also been involved in the adaptation and linguistic verification of the previous version of the test, the CLA, for various countries in the Assessment of Higher Education Learning Outcomes feasibility study (Tremblay et al. 2012, p. 198; on the general approach, see also Ferrari et al. 2013).

German adaptation and test validation team also had experience in the translation of tests, including over 5 years' worth of prior projects in the areas of business and economics.⁴ Thus, team expertise was deemed adequate for the adaptation of assessments (e.g., Arffman 2013; Behr 2012). The steps of the TRAPD process were carried out under time constraints due to practical reasons of research (see Table 12.1). Given the complexity and novelty of the CLA+ assessment, the adaptability and suitability of the test for Germany had to be critically evaluated following each validation step (see Table 12.1). The decision to adapt a second PT came as a result of the expert panels (see Sect. 3.2.1). CAPStAn provided the double translation and reconciliation of the assessment, which were subsequently reviewed by the German team in order to ensure a high level of quality of the German test version. The translated materials included two open-ended PTs on topics of health and sports and the 25 SRQs as well as the detailed item scoring guidelines for the CLA+. CAPStAn translators based their work on experience and general guidelines from previous projects, for instance, from the adaptation of the AHELO study (AHELO 2011) or the Programme for International Student Assessment by the Organisation for Economic Co-operation and Development (OECD) (PISA 2010). Specific problem-oriented translation guidelines for the CLA+, such as documenting all linguistic and cultural translation problems sentence by sentence, sometimes requiring adaptation, were drafted and agreed upon by cApStAn, CAE, and the German team. They were based on guidelines drafted previously for the Italian adaptation of the same CLA+ tasks, which were designed to facilitate cross-national comparisons between Italy and Germany later on.

The translation process itself varied due to the complexity of the items. In addition to 25 SRQs, 2 PTs were selected that were deemed generally adaptable to a German context. The translatability evaluation was supported by the item-specific translation guidelines. The SRQs only presented minor adaptation challenges. For the 1st PT on health, no major cultural differences were identified, which is why it was first to be adapted. In turn, both the analysis of translatability and expert panels (see Sect. 3.2.1) indicated major cultural differences for the 2nd PT on sports. Various aspects of the baseball scenario would have been unfamiliar to students in Germany or implausible in a German context. However, since the experts had judged the test, in particular the 2nd PT (see Sect. 3.2.1), to be generally relevant for higher education in Germany, the German team decided to explore various adaptation strategies. First, as was the case with the other parts of the CLA+,⁵ the 2nd PT was translated by two translators independently, and the preliminary versions were reconciled by a senior translator at cApStAn. Necessary cultural adaptations were documented beforehand and discussed between test validators and translators. The

⁴For example, on the adaptation of the Test of Understanding in College Economics (TUCE) and the Examen General de Egreso de la Licenciatura (EGEL) in the WiWiKom project, see Brückner et al. (2014).

⁵The scoring guidelines were translated by one translator only, as they would be rephrased by the test validators in Germany in line with the German conceptualization of the construct, as advised by CAE.

initial assignment for the 2nd PT was to preserve the original baseball context and adapt it as little as possible. This translation strategy, discussed in survey translation under the term *ask-the-same-question* approach (Mohler et al. 2016), aims to alter the original item composition as little as possible in order to preserve psychometric properties (across several languages), but also bears the risk that students might consider the item “foreign” or difficult to understand. The interviewed experts (see Sect. 3.2.1) concerned that German students might have difficulty picturing themselves as part of a group of decision makers in the United States and suggested rather to prompt them to assume the role of foreign advisors as opposed to decision makers in the United States. This adaptation would have affected only a small part of the text, but the consequences for test performance and cross-national comparability would have been difficult to foresee. Instead of selecting one alternative, the German team ultimately decided to test the effects of a nonadapted version against a localized version.

As a consequence, an *ask-a-different-question* (Mohler et al. 2016) approach was applied to produce a second, fully localized version of the same PT. To this end, previous work materials including the first translation and translation guidelines were used as input, and the entire TRAPD process was reapplied from the start. To control and better document the production conditions of this localized version for subsequent research, this second version was translated and localized entirely at the Faculty of Translation Studies of Mainz University. Starting with the translation of the scoring guidelines, the team identified major lines of reasoning and the supporting dimensions of meaning in the scenario context. Then, an assessment of translatability was carried out, which identified general translation problems, also reflected in cApStAn’s specific scoring guidelines. The localization of realistic micro case studies in the PT was particularly challenging and required in-depth research in order to find German equivalents. In this, the scoring guidelines were helpful for preserving the most relevant item aspects. Various alternatives that covered the same dimensions of the domain of sports in Germany and the United States were discussed. Based on the decision to place the popular sport of soccer at the center of the German scenario, the rest of the task was localized, while the overall structure of the item and approximate amount of distractor information were maintained. In addition to the adaptation of the item text, the localization of graphics was also recommended, both for cultural reasons and for matching the information in the text. This work will require further testing in cross-national comparability analyses (e.g., of effects of cross-cultural differences in illustrations, see Solano-Flores et al. 2016).

The localized version is currently being validated for future use in assessment in Germany (see Table 12.1). The localization illustrates the generally interpretative nature of the translation and adaptation process and the need for close cooperation between test developers and translators. Correspondingly, an additional review based on a back translation is being carried out by CAE for further quality assurance. Other quality assurance measures included, for example, terminology management to ensure consistency within and across tasks and proofreading by two professional translators to ensure linguistic quality. Overall, the translation process complied with the highest academic quality standards.

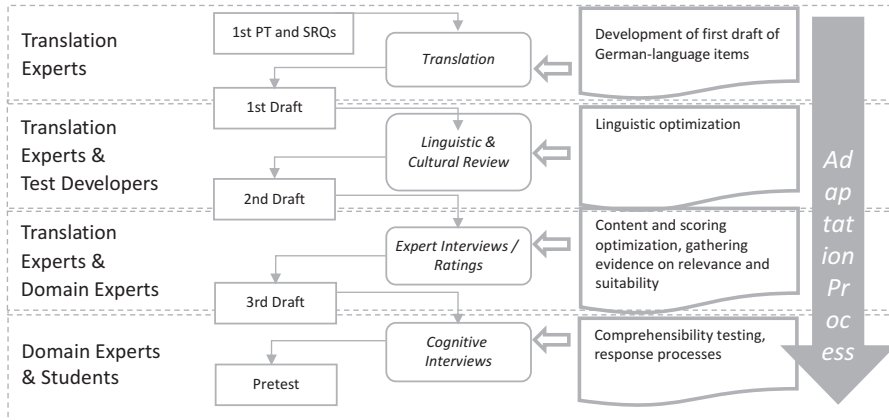


Fig. 12.1 Translation, adaptation, and validation process for the 1st PT in the German study

The extensive validation procedures (see Sects. 3.2 and 3.3) served to continue to systematically enhance certain aspects of the items. Specific translation problems were discussed in a workshop with experts from various areas. In several feedback sessions, experts reviewed the test and reported shortcomings. The translated versions were revised in further workshops with experts in translation studies (corresponding to the step of Adjudication). Cognitive interviews with students offered indications on whether the items contained any remaining passages that were difficult to understand or unintentionally misleading (see Fig. 12.1). A first version of the CLA+, including 2 PTs and the 25 SRQs, was successfully adapted for use in Germany. Pending successful validation, the localized 2nd PT will be examined further to enhance the quality of adaptations.

12.3.2 Preliminary Findings from the Test Validation

12.3.2.1 Expert Panels

In order to validate the construct underlying the two PTs and SRQs with the validation criterion (1) of the Standards (AERA et al. 2014), three expert workshops were carried out between December 2015 and February 2016 in Mainz and Berlin. The first two workshops aimed to evaluate the partially adapted instrument in terms of its general suitability, content validity, and curricular relevance for use in higher education in Germany. The first panel (December 2015, Mainz) focused on the content-related validation and relevance to the curriculum of the 1st PT *life expectancy* (PT1) and the SRQs. In the second panel (January 2016, Berlin), the construct definition and its operationalization in the two PTs and SRQs were critically discussed with experts on psychometrics and experts on the assessment of higher-order cognitive skills, such as problem solving. In the first and second workshops,

respectively, ten experts from various subject areas (including lecturers in biology, business and economics, chemistry, English linguistics and translation studies, higher education research, mathematics, medicine, physics, psychology, and social and political sciences) from different German universities discussed item quality, domain-specificity and generality, challenges of transdisciplinary, cross-institutional, and cross-national testing and comparative analyses, scoring problems, necessary additions, and optional modifications.

The third panel (February 2016, Mainz) focused on the content validation of the 2nd PT *stadium building* (PT2) as well as the evaluation of the translation of both PTs and SRQs. Together with three experts from the field of linguistics and translation studies, the items were discussed with regard to their acceptability and need for further adaptation for Germany to achieve the project aims.

All three expert panels took place in the form of topic-focused, structured group discussions, which were recorded and examined through content analysis. The results of each panel formed the basis for further adaptation and validation (see Table 12.1). For example, findings on the SRQs fed into the subsequent online rating by experts (see below).

First Results

1. *Construct definition.* In all workshops, experts recommended that the construct to be assessed should be defined more clearly for Germany and linked to theory and empirical data. The individual dimensions of the construct should be substantiated a posteriori. During the discussion, culture-specific particularities and cross-national differences in central aspects of the construct and terminology, such as critical thinking and problem solving, became evident; they were attributed to different scientific traditions and a different understanding of academia in the United States compared to Europe or Germany. For instance, German experts were concerned that the two PTs would assess different dimensions of critical thinking. The 1st PT would assess the ability to “deal with (scientific) evidence,” “evidence-based argumentation,” or the “competency to evaluate information,” whereas the 2nd PT would rather assess “problem solving.” For an additional specification of the construct and test definition, it was suggested to link the dimensions examined in the construct to categories of scientific theory or philosophy (e.g., analytical-logical argumentation), in order to specify hypotheses and differentiate the scoring more precisely on this basis.

2. *Further development of the adapted test instrument.* All experts suggested that in order to further develop the instrument and assess important facets of critical thinking, further questions should be added to the tasks. These questions should ask students to evaluate whether they need additional information to solve the task or whether some of the information given was unnecessary and to rate the quality and credibility of the information sources and evidence. All experts pointed out that it would be indispensable to critically examine the extent to which the assessed skills in fact correspond with academically taught competencies. With regard to potential construct-relevant influence factors, the experts identified prior knowledge and skills such as the ability to read diagrams that can determine performance on the

item. Such skills are mainly acquired at school in Germany; hence, greater attention should be given to assessing students' preconditions. Therefore, all experts recommended assessing and controlling for additional individual student characteristics and influence factors in subsequent validation analyses, including controlling for reading comprehension, intelligence, language skills, ability of abstraction, and attitudes or epistemological beliefs.

3. *Relevance to everyday life and sensitivity to study domain.* All experts pointed out the practical relevance of the PTs as particular strength of this assessment. There were, however, critical discussions about the extent to which the instrument assesses generic abilities or rather subject-specific skills acquired in higher education. The experts unanimously pointed out that the instrument might not be suitable for comparisons across disciplines due to the subject-sensitivity of German higher education. It was criticized that students from certain disciplines in Germany, such as "degree courses without an empirical focus" or "arts degree courses," would have a disadvantage in the test, whereas, for example, medical students or students of life sciences were expected to achieve better results on the 1st PT.

4. *Cultural and linguistic comparability of the adapted instrument.* The experts discussed whether the original CLA+ instrument and its adaptations (so far in England, Germany, and Italy, Zahner and Cioffi, Chap. 11 in this volume) could be generally suitable for an international comparative study. Challenges of linguistic and cultural comparability were identified in particular for the PTs and their scoring. The experts questioned whether the scoring criteria *writing effectiveness* and *writing mechanics* could be compared across countries. While the scenario of the 1st PT was judged to be understandable for German addressees without major adaptations and therefore cross-culturally comparable, the baseball context of the 2nd PT was judged to be much less typical for the German culture. Comprehension and response processes for the 2nd PT were therefore judged to be more difficult than in the US original. Thus, adaptations proved to be inevitable, even though they might negatively impact measurement equivalence across countries. The adapted task would need to be examined more thoroughly regarding its suitability for an international study (see Sect. 3.1).

Micro adaptations of individual aspects or macro adjustments to the entire text were discussed as possible solutions. On the one hand, the original US context could be maintained with only minor changes; however, in order to make the scenario plausible to students in Germany, they would be prompted to assume the role of external, international consultants for another country rather than local decision makers. On the other hand, a localized alternative was deemed suitable for higher education in Germany; this would, however, require a comprehensive change of the scenario context, for instance, from baseball to soccer. As pointed out by the experts, this option would involve risks of altering the psychometric properties of the item, affecting subsequent international comparisons.

5. *Equivalence of different performance tasks and scoring.* With regard to the potential parallel use of both PTs in a field study, the experts compared underlying construct definitions and relations to domains and culture. As noted above, the PTs

were judged to measure different facets of critical thinking and could therefore not be used for comparisons without further analyses. Furthermore, the experts estimated that participants' individual motivation and interests, such as attitudes toward healthy living or particular interests in sports, could confound performance on the items. Possible cultural or gender effects were also expected due to the scenario contexts. According to the experts, the two different PTs allow assessment of different construct facets in higher education, such as a critical approach to sources and evidence, argumentation, or problem solving. Similar questions were raised for the scoring, which was judged problematic when used as a uniform scoring across PTs. Suggestions were made such as giving up the holistic coding scheme, designing more differentiated scoring categories, optimizing the fit between scoring and item instructions in the German version, and using experimental responses from the validation studies to further develop the scoring. The categories could also be defined based on the facets of the German construct definition. In this case, however, international comparability of the scoring might be problematic.

Overall, the expert panels indicated that the CLA+ is an innovative approach to performance assessment that is relevant for higher education practice; the assessment format was judged an interesting and useful addition to current examination practice in Germany. However, experts recognized various challenges to be addressed before the instrument could be used in Germany as well as in an international study, including appropriately adapting the instrument for the higher education context in Germany. This concerns questions of domain-specificity of scenarios and dependence of student performance on prior subject knowledge, which would make it more difficult to use the instrument across disciplines and institutes. It also refers to the equivalence of the construct, dimensions, and facets assessed by the two PTs. Moreover, experts critically discussed the extent to which the test assesses skills acquired in higher education rather than preconditions acquired in upper secondary education in Germany. In accordance with the construct, which needs further differentiation, revisions should be made to the scoring, which should be more closely aligned to the facets of the construct definition and could be developed on the basis of the experimental responses. Further insights to guide necessary modifications were expected from the cognitive interviews, which, according to the experts, were a suitable approach for validating comprehension of and mental response processes to the two adapted PTs.

12.3.2.2 Curricular Analysis and SRQ Rating

In a preliminary curricular analysis, examining whether generic skills in general and the test content of the CLA+ in particular represents part of the curriculum in various fields of studies in higher education in Germany, curricula and module descriptions from 32 different degree courses were analyzed. Overall, the curricular analyses suggested that the adapted item content of the CLA+ is part of curricula in higher education in Germany. In addition, curricular relevance and content validity

were supported by the experts' evaluations during the expert workshops and online expert rating, which indicated that these types of skills assessed are being taught at higher education institutions in Germany.

The SRQs were rated by 12 professors and lecturers at higher education institutions in Germany. This expert rating served to cross-validate the curricular analyses and to evaluate additional aspects that were relevant to content validation. The experts rated the curricular relevance and the difficulty of the items and gave a general evaluation of each item. To keep the experts' work within acceptable limits, each of them was asked to rate no more than four items. The questionnaire included closed-ended rating items on a seven-point Likert scale as well as open questions and feedback areas for general concluding remarks.⁶ All experts rated both the difficulty and the complexity of the test tasks as appropriate for undergraduate students across the different fields of studies. Additionally, the question of whether the test tasks capture central facets of generic skills relevant to the higher education has also been judged as appropriate by the experts. In particular, the experts regarded the relevance of the test facets for the transition to the job market as strong. Overall, content validity was confirmed for all adapted SRQ items from the CLA+. The findings also suggested that the constructs of generic skills were understood in a similar way in different study domains at various universities (for more details, see Kaufmann 2017).

Content validation was interlinked partly with the adaptation (see Sect. 3.1) and was followed by cognitive interviews.

12.3.3 Cognitive Interviews As a Validation Measure

For the validation of the translated, linguistically and culturally adapted PT1, as well as of the translated and linguistically adapted PT2,⁷ cognitive interviews were conducted with ten students, drawn by a purposeful sampling (Miles and Huberman 1994) to explore their understanding of the items as well as to identify and analyze mental processes occurring during the response process. The sample included beginner and advanced students, students from different study domains and from different performance levels in order to allow for the observation of possible effects of different domain-specific contexts and learning experiences versus generic skills and attitudes⁸ when solving the PT.

⁶For example, "Does the item represent a higher education curriculum or a higher education domain?" "In what ways are constructs likely to differ across German higher education institutions?"

⁷Because of the specific content and context, the cultural adaptation of the PT2 was initially forgone. A cultural adaptation of the PT2 was conducted at Faculty 06 of Mainz University in the summer semester of 2016. Further coglabs have been conducted on both the culturally adapted and the nonculturally adapted version of PT2.

⁸For example, the sample included one student from the domain of medicine who was particularly interested in a healthy lifestyle.

12.3.3.1 Aim of the Cognitive Interviews

Cognitive interviews are used in a multitude of areas in test development and validation. They assess not only formal aspects such as comprehensibility and correctness in the phrasing of tasks but also more complex aspects of a process-related analysis of the mental processes during task-solving in order to derive significant insights about the assessed construct, especially with regard to cognitive validation (Brückner and Pellegrino 2016; Leighton 2013). Another field of application is the linguistic and cultural adaptation of test instruments as well as translation research (e.g., Willis 2005; Fitzgerald et al. 2011; Goerman 2006; on the cognitive validations of CLA tasks in the context of the AHELO study, see Hyytinen et al. 2014).

A cognitive interview study preceded the field application as a pretest, aiming to create functionally equivalent tasks for multiple languages in which CLA+ is used. In cooperation with researchers from different domains (e.g., economists, translation experts, and psychologists), the tasks on life expectancy and the building of a stadium were adapted from the US – American context for the German linguistic and cultural background (see Sect. 3.1). Then, the tasks were assessed in cognitive interviews with regard to their alignment with the understanding of test developers: “These techniques are used to examine whether respondents’ interpretations of [self-report] items are consistent with researchers’ assumptions and intended meanings given the constructs the items are designed to measure” (Karabenick et al. 2007, p. 139).

The intention to analyze the equivalence between the two tasks and the related mental processes justified by the fact that the tasks were developed from different linguistic and cultural contexts which potentially have a divergent understanding of certain concepts and can therefore present culture-specific peculiarities which may need to be adapted (see Sect. 3.1). An excellent example is the original PT2 from the American context, which is about the building of a baseball stadium. In Germany, however, baseball is not a popular sport; therefore, German students might have more difficulties solving this task, as they can hardly comprehend the cultural and contextual significance of building such a stadium in Germany. Here, the question ensues whether the task should be adapted for the German context in building a new, for example, soccer stadium.

The benefits of the think-aloud methods have been “rediscovered” over the last few years (e.g., Leighton 2013) in order to enable a comparison of measuring instruments from different linguistic and cultural contexts based on mental processes (Goerman 2006). The German study also used this method and embedded it in an assessment design in order to evaluate comparability and to be compatible with the pretest procedures with the CLA+ from previous adaptation processes in other countries (see also Zahner and Ciofi, Chap. 11 in this volume, Solano-Flores et al. n.d.).

12.3.3.2 Preparation and Conduction of the Interviews

Overall, ten students from different degree courses (economics, education, medicine, cultural studies, sociology, politics) were interviewed, six of whom were given the PT on life expectancy, and four were given the PT on the building of the stadium. The interviews were conducted according to a standardized procedure (Solano-Flores et al. n.D.).⁹

Before the beginning of each interview, students were told that the aim was not to test them but to assess the adapted German test versions. As the task documents include a lot of graphs and tables, an intelligence test (IST, Liepmann et al. 2007) with visual tasks was conducted with each student. Then, they were subjected to a short training on thinking aloud. Student could voice potential reservations to receive clarification. After giving a method description, training the students with simple “warm-up” exercises, and asking them to confirm their understanding and ability to think aloud, test coordinators conducted the actual thinking aloud interviews. At the end of each interview, some socio-biographical data were gathered, as well (e.g., degree course, gender, study progress).

Before both the *concurrent* and the *retrospective* interview phase, students were once again explained the purpose of the interview.¹⁰ The characteristic feature of the concurrent phase was that the students worked on the tasks autonomously and without interacting with the test coordinator; the only interaction were reminders to keep talking when they forgot to say their thoughts aloud for a longer period of time (approx. 10 s). During this phase, the interviewer took notes about, for example, how often the student read a certain sentence or passage repeated or underlined words had difficulties with certain terms. In the second phase, the *retrospective phase*, the test coordinator was allowed to ask the students further questions. In addition, similar to cognitive interviews in ANVUR (Solano-Flores et al. n.d.), in this final phase, a standardized interview guideline was used by the test coordinator to ask 10 questions on different aspects of the tasks and the solving process (see Table 12.2).

The data from both phases will then be discussed with the test developers of the US tasks and compared to the data generated from cognitive interviews with the original English instrument. The comparison will allow for a first insight into the response processes in both countries and indicate need for adaptation.

⁹Test coordinators avoided creating a testing atmosphere by seating themselves inclined to the assesse, positioning video recording devices out of sight, and maintaining a disturbance-free environment. In addition, data privacy was observed by filming only the respondents’ hands and multiple test documents.

¹⁰The note they were read said: “With this interview, we want to investigate how students handle information that they come across in everyday life. For this purpose, we developed a test and we now want to find out whether the tasks that we developed are suitable for use in higher education. It is therefore not the aim of this experiment to measure your expertise; the results will have no influence on your grades whatsoever. We are interested in how students handle the task, how they solve it and what thoughts cross their minds in the process. We would therefore like to ask you to say everything you are thinking out loud while working on the task, even when you have an idea and then end up dismissing it or when you seem to not understand a word! Everything you would say silently to yourself, you should please say out loud. Just imagine you are alone in the room.”

Table 12.2 Standardized questions of the retrospective phase

Coglab questionnaire
Please summarize how you arrived at your solution.
What information did you find to be especially helpful in responding to the item?
Under which circumstances would you have perhaps argued differently?
What did you find especially difficult about the task?
Which materials or information would you have needed in order to solve the task in a satisfactory way?
How did you decide which information is especially relevant for you to solve the task?
Which strategy did you use to respond to the task?
Did you find the task motivating? If yes, why? If no, why not?
How realistic do you consider the situation described in the task?
To what extent do you think the tasks could help you prepare for a professional career?

12.3.3.3 Preliminary Results

With a range from 18 to 29 years the average age of the participants was 23.2 years. Two thirds of the students were female and one third was male students. While the sample showed differences in the family background of the participants – 22.2% indicated that at least one parent originates from another country than Germany and the educational qualification of the parents ranged from a high school diploma to a doctoral degree – all participants stated that the most commonly spoken language in their family environment was German. The sample did also vary regarding the grade on the higher education entrance qualification: a variation from 1.6 to 3.3 could be determined with an average of 2.5.

All students were asked to fill in self-evaluations, which contained four questions. The first two questions concerned the possible disruptions through thinking aloud and the presence of the interviewer. The disruption through thinking aloud was experienced differently by the students – with a mean score of 2.56 on a scale from 1 (not at all) to 5 (a lot). In comparison, all of the students stated that they were “not at all” (1) or “a little” (2) disrupted by the presence of the test coordinator. The third question asked about the interviewer’s expertise regarding critical thinking, which was answered with an average of 4.0 on a scale of 1 (very low) to 5 (very high). Through the last question, concerning the willingness of the students to participate in the study, it was shown that the participants were highly motivated (average 4.22).

The results of the figural and verbal analogies IQ tests conducted with each student revealed large differences between the students – figural test: min. 4 and max. 19 right answers out of 20 tasks; test about analogies: min. 2 and max. 16 right answers out of 20 tasks. While male participants performed better on both IQ tests – figural test: male average 14.67 and female average 10.5; analogy test: male average 13.33 and female average 8.67 – test results also showed correlations with parents’ origin and the grade of the higher education entrance qualification.

Further findings indicate that the time of item responding varies between students. Some students needed merely 40 min to solve a PT, while others needed nearly twice as much time. A large part of solving time was spent for studying the provided documents. Typically, a student who solved the tasks in 60 min initially spent nearly 30 min reading and understanding the documents, 2 min for reading the task description, 13 min for rereading the documents and selecting and noting down the most important pieces of information and arguments, and 15 min for finally writing down the answer. Generally, however, all students believed that the target solving time of 60 min should be increased by approximately 20 min.

In terms of content, we observed that many students perceived the topics of the tasks as interesting but were not necessarily motivated to process and solve them. This overlaps with the experiences made by the test developers in the United States, who also reported motivational limitations in item responding. The problem situation described in both tasks was perceived as realistic by many students, even though in the life expectancy task they would have liked to have had more information on the topics of exercise and sleep instead of nutrition and diet. Such information seemed helpful to them as a multifactorial construct. The relations to everyday real life also became evident, as many students perceived the tasks to be useful in preparation for their future professional life. For example, it was pointed out that solving the task helped to use information presented through various media more critically. Furthermore, it was noted that in one's life, both professional and private, one is repeatedly confronted with decisions and that it is therefore helpful to learn to weigh different arguments against one another. However, in order to create an even higher relevance to future professional activities, the students would have liked different, more (domain) specific contents so that the task would specifically prepare them for their professional life.

12.4 Conclusion and Outlook

In this study, we adapted and validated the internationally proven performance assessment CLA+ for Germany, taking into account the underlying conceptual model and assessment framework. For this purpose, we took a multi-perspective and multi-method qualitative approach in examining, among others, the content validity and curricular relevance of the assessment for higher education in Germany as well as the underlying response processes and mental operations. By further in-depth analyses of the think-aloud protocols, we will be able to explore whether item responses of different groups of students were based on different mental processes and representations or different test-taking strategies.

The preliminary results from our validation study showed that this performance assessment enables measuring higher-order cognitive skills at the academic level in higher education. This kind of assessment is innovative for higher education practice in Germany and has significant potential for enhancing curricula and instruction to promote students' interdisciplinary skills. Yet, further research and development are needed in particular with a focus on the concept and test definition. The question

as to which concrete skills are assessed with the PTs and SRQs remains unclear and requires further theoretical and empirical research. Another issue lies with the further examination of domain-specificity and the extent to which generic skills can be assessed through specific situational contents and contexts which make reference to certain domains. In other words, the question is whether the same skills can be assessed despite different contents of the tasks.

When implementing this kind of assessment, a number of practical issues arise, such as the question of ensuring test security and test motivation. Our preliminary results show that test motivation is very strongly dependent on the students' interest in the item context, for example, in a healthy lifestyle or a certain sport. Overall, many of the interviewed students would have liked to see a stronger connection to their respective study domains to find the tasks more interesting, which would be problematic with regard to domain-specificity. Should it be possible to assess the same skills using different contents and contexts, it would be possible to let students choose from a pool of tasks. To this end, however, further analyses of the internal test structure are necessary for Germany to empirically prove that all tasks within the item pool assess the same skills and that the test results are comparable. The expert interviews and discussions with professors and lecturers indicated that the implementation of such assessments in higher education practice should be accompanied by corresponding teaching and learning tools. For the United States, CAE has already developed such a tool and reported positive experiences.

To what extent this assessment is suitable for intra- or cross-institutional comparisons remains to be explored in further research. This also holds true for comparisons with other countries. To this end, different adapted versions shall be examined with regard to their measurement equivalence in order to ensure that the adapted tasks measure the same skills and to determine which further adaptations are necessary. Conducting cognitive labs on all adapted versions would be desirable in order to explore whether the same cognitive thought operations are used for responding to adapted versions. Another useful complementation would be to conduct eye-tracking studies in order to control, for example, the effects of general reading abilities, such as reading speed.

References

- AHELO Consortium. (2011). *Translation and adaptation manual/guidelines*.
- 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.
- Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educational Measurement: Issues & Practice*, 32(2), 2–14.
- Behr, D. (2012). The team translation approach in questionnaire translation: a special form of expert collaboration. In *Proceedings of the 2nd international specialist conference of the German Federal Association of Interpreters and Translators (BDÜ) 28–30 September 2012* (pp. 644–651). BDÜ, 32. Berlin: BDÜ.

- Behr, D., & Shishido, K. (2016). The translation of measurement instruments for cross-cultural surveys. In C. Wolf, D. Joye, T. W. Smith, & Y.-C. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 269–287). London: Sage.
- Braun, M. (2006). *Funktionale Äquivalenz in interkulturell vergleichenden Umfragen. Mythos und Realität* [Functional equivalence in comparative intercultural surveys: myth and reality.] Mannheim: ZUMA.
- Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multi-level models to validate an assessment of higher education students' competency in business and economics. *Journal of Educational Measurement*, 53(3), 293–312.
- Brückner, S., Zlatkin-Troitschanskaia, O., & Förster, M. (2014). Relevance of adaptation and validation for international comparative research on competencies in higher education – A methodological overview and example from an international comparative project within the KoKoHs research program. In F. Musekamp & G. Spöttl (Eds.), *Competence in higher education and the working environment. National and international approaches for assessing engineering competence, Vocational education and training: Research and practice* (Vol. 12, pp. 133–152). Frankfurt am Main: Lang.
- Coates, H. (Ed.). (2014). *Higher education learning outcomes assessment – International perspectives*. Frankfurt/Main: Peter Lang.
- Council for Aid to Education (CAE). (2013). *Introducing CLA+ Fostering great critical thinkers*. New York: CAE. http://cae.org/images/uploads/pdf/Introduction_to_CLA_Plus.pdf
- Council for Aid to Education (CAE). (2015). *The case for generic skills and performance assessment in the United States and international settings*. New York: Council for Aid to Education. http://cae.org/images/uploads/pdf/The_Case_for_Generic_Skills_and_Performance_Assessment.pdf
- Coyne, I. (Ed.). (2000). *International test commission test adaptation guidelines*. Accessed 11 December from: www.intestcom.org/test_adaptation
- European Commission (EC). (2015). *European qualifications framework*. https://ec.europa.eu/ploteus/search/site?f%5B0%5D=im_field_entity_type%3A97
- Ferrari, A., Wayrynen, L., Behr, D., & Zabal, A. (2013). Translation, adaptation, and verification of test and survey materials. In OECD *Technical report of the survey of adult skills (PIAAC) 2013* (pp. 1–28, section 1, chapter 4). http://www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf. Accessed 14 Jan 2017
- Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2011). Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. *Journal of Official Statistics*, 27(4), 569–599.
- Förster, M. (2004). Higher order thinking skills. *Research Developments*, 11, article 1. <http://research.acer.edu.au/resdev/vol11/iss11/1>
- Förster, M., Happ, R., & Molerov, D. (2017). Using the U.S. test of financial literacy in Germany – Adaptation and validation. *The Journal of Economic Education*, 48(2), 123.
- Goerman, P. L. (2006). An examination of pretesting methods for multicultural, multilingual surveys: The use of cognitive interviews to test Spanish instruments. In J Harkness (Ed.), *GESIS-ZUMA (Ed.): Conducting cross-national and cross-cultural surveys: papers from the 2005 meeting of the international workshop on Comparative Survey Design and Implementation (CSDI)*. Mannheim, 2006 (ZUMA-12).
- Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, 17(3), 164–172.
- Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken, NJ: Wiley.
- Herl, H. E., O'Neil, H. F., Jr., Chung, G. K. W. K., Dennis, R. A., Klein, D. C. D., Schacter, J., & Baker, E. L. (1996). *Measurement of learning across five areas of cognitive competency: Design of an integrated simulation approach to measurement. Year 1 report*. Los Angeles: University of California.

- Hyytinen, H., Holma, K., Toom, A., Shavelson, R. J., & Lindblom-Ylänne, S. (2014). The complex relationship between students' critical thinking and epistemological beliefs in the context of problem solving. *Frontline Learning Research*, 2(5), 1–25.
- International Test Commission (ITC). (2016). *The ITC guidelines for translating and adapting tests* (2nd ed.). www.InTestCom.org. Accessed 14 Jan 2017.
- Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, V. B., Blazevksi, J., Bonney, C. R., et al. (2007). Cognitive processing of self report items in educational research: Do they think what we mean? *Educational Psychologist*, 42(3), 139–151.
- Kaufmann, F. (2017). *Validierung des Testinhalts eines Kompetenzerfassungsinstruments anhand von Expertenratings*. Unveröffentlichte Masterarbeit.
- Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education*, 26(2), 136–157.
- Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R* [Intelligence Structure Test] (2., erweiterte und überarbeitete Aufl.). Göttingen: Hogrefe & Huber Publishers.
- Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next-generation assessment* (Research Report). Princeton: ETS.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Mohler, P., Dorer, B., de Jong, J., & Hu, M. (2016). *Translation: Overview. Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.
- Organisation for Economic Co-operation and Development (OECD). (2013). *Assessment of higher education learning outcomes. AHELO feasibility study report – Volume 2. Data analysis and national experiences*. Paris: OECD.
- Organisation for Economic Co-operation and Development (OECD). (2014). *Education at a glance 2014: OECD indicators*. Paris: OECD Publishing. <https://doi.org/10.1787/eag-2014-en>
- Organisation for Economic Co-operation and Development (OECD). (2016). *Getting skills right: Sweden*. Paris: OECD Publishing. <https://doi.org/10.1787/9789264265479-en>
- Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.
- PISA Consortium. (2010). *Translation and adaptation guidelines for PISA 2012*. National Project Managers' meeting, Budapest 2010. <https://www.oecd.org/pisa/pisaproducts/49273486.pdf>. Accessed 14 Jan 2017.
- Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33(6), 569–600.
- Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist*, 48(2), 73–86.
- Shavelson, R. J., Davey, T., Ferrara, S., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Princeton, NJ: Educational Testing Service.
- Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 45–60). Hillsdale, NJ: Erlbaum.
- Solano-Flores, G., Wang, C., & Shade, C. (2016). International semiotics: Item difficulty and the complexity of science item illustrations in the PISA-2009 international test comparison. *International Journal of Testing*, 16(3), 205.
- Solano-Flores, G., Chia, M., Shavelson, R. J., & Kurpius, A., (n.d.). *CAE cognitive labs guidelines*. Unpublished document by the Council for Aid to Education. New York

- Tremblay, K. (2013). OECD assessment of higher education learning outcomes (AHELO): Rationale, challenges and initial insights from the feasibility study. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education* (pp. 113–116). Rotterdam: Sense Publishers.
- Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education learning outcomes. Feasibility study report. Volume 1 – Design and implementation*. OECD. <http://www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf>
- Wheeler, P., & Haertel, G. (1993). *Resource handbook on performance assessment and measurement*. Berkeley, CA: The Owl Press.
- Willis, G. B. (2005). *Cognitive interviewing. A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.
- 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: Peter Lang.
- Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, 40(3), 393–411.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Lautenbach, C., & Toepper, M. (2016a). Assessment practices in higher education and results of the German research program modeling and measuring competencies in higher education (KoKoHs). *Journal Research & Practice in Assessment*, 11, 46–54.
- Zlatkin-Troitschanskaia, O., Pant, H. A., Kuhn, C., Toepper, M., & Lautenbach, C. (2016b). *Messung akademisch vermittelter Kompetenzen von Studierenden und Hochschulabsolventen. Ein Überblick zum nationalen und internationalen Forschungsstand* [Assessment of academic competencies of students and graduates – An overview of the national and international state of research]. Wiesbaden: Springer
- 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.

Chapter 13

Validating the Use of Translated and Adapted *HEIghten*[®] Quantitative Literacy Test in Russia



Lin Gu, Ou Lydia Liu, Jun Xu, Elena Kardonova, Igor Chirikov, Guirong Li, Shangfeng Hu, Ningning Yu, Liping Ma, Fei Guo, Qi Su, Jinghuan Shi, Henry Shi, and Prashant Loyalka

Abstract In responding to the need for internationally comparable data on higher education student learning outcomes, some modules of the *HEIghten*[®] Outcomes Assessment Suite, developed by Educational Testing Service, have been translated and adapted for international use. This recent development points to a critical need to validate the use of translated and adapted *HEIghten* assessments in international contexts. This chapter reports on validating the use of the Russian *HEIghten* Quantitative Literacy (QL) assessment with a representative group of students majoring in electrical engineering and computer science from 34 higher education institutions in Russia. Our findings provided preliminary evidence in support of the use of the assessment for the target population as a measure of QL. Future research is suggested to further investigate the test's ability of reflecting changes in the target construct as a function of learning in the context of Russia.

L. Gu (✉) · O. L. Liu · J. Xu
Educational Testing Service, 660 Rosedale Road MS 14-R, Princeton, NJ 08541, USA
e-mail: lgu001@ets.org; lliu@ets.org; jxu@ets.org

E. Kardonova · I. Chirikov
National Research University Higher School of Economics, Moscow, Russia
e-mail: e_kardonova@mail.ru

G. Li
Henan University, Kaifeng, China

S. Hu
Sichuan Normal University, Chengdu, China
e-mail: husf1999@163.com

N. Yu
Shandong Jinan University, Jinan, China
e-mail: yuning7412@126.com

13.1 Introduction

One of the global trends that has been reshaping the landscape of the higher education sector is greater internationalization (e.g., Tremblay et al. 2012). As defined by Knight (2003), internationalization in the context of higher education refers to the “process of integrating an international, intercultural or global dimension into the purpose, functions or delivery of post-secondary education” (p. 2). Taking various forms, from globalized curricula to international mobility of students and academic staff, internationalization is playing an increasingly critical role in determining and influencing national and institutional strategy and policy (Organisation for Economic Co-operation and Development (OECD) 2008). A major driving force behind this trend is the increasingly interconnected world economy, which demands a labor force equipped with skills that are important for operating successfully on a global scale (e.g., Bennell and Pierce 2003; see also Coates, Chap. 1 in this volume).

Against this backdrop of greater internationalization in higher education motivated by an interconnected world economy lies an emerging need for internationally comparable information and data on student learning outcomes (SLOs, Tremblay et al. 2012). Performance data based on comparative assessments as well as contextual information associated with the performance are relevant to a variety of stakeholders, including governments, higher education institutions, and student learners, and can be used to establish international standards and benchmarks against which SLOs can be evaluated in a comparative framework (see also Shavelson et al., Chap. 10 in this volume). Information gained through such a comparative perspective can be expected to yield immediate effects, for example, informing national strategies and policies at the national level, facilitating internal improvements at the institutional level, and enabling skill diagnosis at the student level. These immediate effects are then expected promote long-term effects on learning, that is, to help facilitate the acquisition of twenty-first century skills which can then empower individuals to seek greater social mobility, contributing to greater social equality (cf. Alexander, Chap. 3 in this volume).

L. Ma

Peking University, Beijing, China
e-mail: lpma@pku.edu.cn

F. Guo · J. Shi

Tsinghua University, Beijing, China
e-mail: feigu0121@mail.tsinghua.edu.cn; shijhuan@tsinghua.edu.cn

Q. Su

Shaanxi Normal University, Xi'an, China
e-mail: suqiceee@163.com

H. Shi · P. Loyalka

Stanford University, Stanford, CA, USA
e-mail: loyalka@stanford.edu

Some indicators of higher education quality have been used to gain comparative insights, including student engagement and satisfaction surveys, university rankings, and employment-based labor market outcomes (e.g., employer feedback, salaries). However, none of the existing approaches provide direct evidence of learning outcomes. Consequently, the absence of assessment tools for evaluating learning outcomes directly on an international scale prohibits the establishment of objective international benchmarks by which the quality of higher education can be evaluated in a comparative framework. In sum, there is a critical need for a comparative assessment of higher education learning outcomes.

In response to the need for accreditation and curriculum improvement by higher education institutions in the US, the *HEIghten*[®] Outcomes Assessment Suite, developed by Educational Testing Service (ETS), measures the skills and competencies deemed critical for both higher education and the workforce. Through research initiatives with institutions across the globe, some of the *HEIghten* assessment modules have been translated and adapted from the original source language (i.e., English) to different languages to be used for measuring SLOs in diverse international contexts. These recent developments lay open the possibility for *HEIghten* to be used as a common metric for facilitating international comparisons of SLOs.

A prerequisite condition that needs to be satisfied for making meaningful cross-country comparisons using *HEIghten* assessments is that scores based on the translated and adapted *HEIghten* assessments are fair, meaningful, and valid for intended uses in various international contexts. According to the International Test Commission (ITC) Guidelines for Translating and Adapting Tests (International Test Commission 2005), information on the evaluation of validity should be provided for all target populations for whom the adapted versions are intended. Based on the framework proposed by the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) 2014; hereafter referred to as *Standards*), it is critical to examine validity evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing. For translated and adapted tests intended for international use, a few validity concerns deserve special attention (see also Zlatkin-Troitschanskaia et al., Chap. 12 in this volume). Regarding evidence based on test content and response processes, of particular concern is the extent to which construct-irrelevant variance introduced by translation/adaptation is minimized. This can be examined by evaluating the relationship between the content of the translated and adapted test and the target construct, as well as the fit between the response processes elicited by translated/adapted test items and the target construct. Concerning evidence based on internal structure, it is important to examine whether the translated/adapted test has appropriate psychometric quality at the item, subscale, and test level in the target population. Critical investigations include test difficulty and discrimination, total scale and subscale reliabilities, differential item functioning, and test dimensionality. Furthermore, the appropriate criteria measures for the translated/adapted test could differ from those of the original test. For example, when *HEIghten* is used in the US context, relevant

external variables include high school/college grade point average (GPA). However, the GPA system might not apply in a foreign context. Therefore, extreme care should be taken to identify the appropriate external variables in the context where the translated/adapted test is used. Last, the sociocultural and ecological contexts of the populations should be taken into consideration and their effects should be accounted for in score interpretation (see also Shavelson et al., Chap. 10 in this volume).

In this chapter, we report on a validation study regarding the use of translated/adapted *HEIghTen* Quantitative Literacy (QL) assessment in Russia. Next, we will introduce the larger research project within which the administration of the Russian *HEIghTen* QL assessment was situated.

13.2 Background of the Larger Research Project

The results reported in this study are part of a larger project led by researchers at Stanford University in collaboration with ETS and researchers from various countries including China and Russia. The overall purpose of this project is to examine learning outcomes for electrical engineering (EE) and computer science (CS) students across multiple countries as well as to help identify which contextual factors impact students' learning. Note that the main purpose of this study is not to rank countries or institutions but rather to gather evidence for improving higher education systems and institutions. To this end, the research team also collected a wealth of contextual survey data from students, faculty, and administrators. To our knowledge, this is the first international comparative project that collected assessment survey data from nationally representative samples of university students majoring in EE and CS.

13.3 Study Purpose and Research Questions

As part of the aforementioned research collaboration, students in Russia took the translated/adapted *HEIghTen* QL assessment. In this study, we aimed to investigate validity evidence regarding the use of the Russian *HEIghTen* QL assessment. We focused on two types of validity evidence, namely, evidence based on the test's internal structure and its relations with external variables by addressing the following two research questions:

- Does the test have appropriate psychometric quality at the item, subscale, and test level for the Russian population?
- Does the test have appropriate relationships with external variables that are construct relevant in the Russian context?

13.4 Data

13.4.1 Sample

In 2015 a total of 1205 Russian college students took one complete form of the *HEIghten* QL assessment that was translated and adapted to Russian. These students were from 34 universities, including 6 elite universities and 28 non-elite universities.¹ Sample size at the university level ranged from 5 to 62 ($M = 34.59$, $SD = 15.28$).

At the time of test-taking, the participants were freshmen (1st-year students) and juniors (3rd-year students), majoring in electrical engineering (EE) or computer science (CS). Motivation screening was applied to remove students who did not complete at least 75% of the assessment.² Using this criterion, 29 students (about 2.4% of the original sample) were removed. Our analysis sample consisted of 1176 students. Table 13.1 summarizes the demographic information of the analysis sample. The majority of the participants came from non-elite schools, accounting for 80% of the sample. The sample was evenly distributed across grade and major (EE or CS). The sample was about 77% male, as gender imbalance is common within STEM (science, technology, engineering, and mathematics) fields. Through an exit survey following the assessment, test takers also provided background information, such as high school type, college entrance exam scores, self-rated QL skills, perceived test difficulty and testing time, etc.

Table 13.1 Demographic information

Demographic information	<i>n</i>	%
Institution type		
Elite	234	19.9
Non-elite	942	80.1
Grade		
Freshmen	656	55.8
Juniors	520	44.2
Major		
CS	668	56.8
EE	508	43.2
Gender		
Male	903	76.8
Female	266	22.6
Other/missing	7	0.6

¹In the Russian higher education system, a university is classified as elite if it has the status of being a Federal University or a National Research University. All six elite universities in our sample were National Research Universities.

²The 75% test completion rate is used for the operational *HEIghten* assessments to screen out students with low testing motivation.

13.4.2 *Russian HEIghten QL Assessment*

The *HEIghten QL* assessment aims to measure the ability to detect and solve mathematical problems in authentic contexts across a variety of mathematical content areas. The assessment framework focuses on two key dimensions: problem-solving skills and mathematical content. The four problem-solving skills measured by the test are (a) interpretation, (b) strategic knowledge and reasoning, (c) modeling, and (d) communication. The four mathematical content areas include (a) number and operations, (b) algebra, (c) geometry and measurement, and (d) statistics and probability. For information on the development of the assessment framework, see Roohr et al. (2014). Validity evidence supporting the use of the assessment in the US context is reported in Roohr et al. (2017).

One complete operational form of the assessment was translated and adapted from English to Russian by cApStAn, a company that provides translation services for large international assessment programs, such as the Programme for International Student Assessment (PISA) and the Programme for International Assessment of Adult Competencies (PIAAC). We adopted the most rigorous, three-step translation/adaptation model that cApStAn offers. First, the test form was double-translated, that is, it was translated by two translators independently. The two translators then reconciled any discrepancies and produced a draft for review. Second, a review team that was fluent in Russian, consisting of content experts and experts in measurement theories and practices from ETS, Stanford University, and Russia, reviewed the draft and provided comments and suggestions for change. Including this broad range of expertise during the review process was considered critical for identifying not only translation problems but also problems that could potentially affect the psychometric quality of the assessment. Following that, cApStAn verified the recommended changes and finalized the translated/adapted test form. The double translation and reconciliation procedure was used by both the PISA and PIAAC programs. This approach was considered to have two significant advantages over back translation, a frequently used translation method, including (a) having multiple people work with both the source and target versions and (b) recording discrepancies directly in the target language instead of in the source language (OECD 2012).

Throughout the translation/adaptation process, particular attention was given to take full account of the linguistic and cultural differences to ensure the comparability of the English and Russian versions of the assessment both at the observed content level and the unobserved construct level. For example, for items that involve currency, where simply changing the currency sign would result in unrealistic numbers in the Russian context, the numerical values were changed to maintain content authenticity. Changing the numerical values in such cases, however, raised the concern about the comparability of the cognitive demands between the original item and the translated/adapted item. To mitigate this potential source for construct non-equivalence, wherever possible, we changed the numerical values by multiplying by powers of ten.

The translated/adapted test was administered online using a research platform that was designed to simulate the operational *HEIghten* testing experience. The test had 25 dichotomously scored items. The total raw score scale ranged from 0 to 25. Table 13.2 shows the number of items by mathematical content and problem-solving skills. Each item targets a problem-solving skill and a content area simultaneously. Note that for the purpose of all subscale analyses reported in this chapter, the two skill sections, communication and interpretation, were combined because the test form had only one communication item.

Table 13.2 Number of test items by sub-construct area

	Sub-construct	<i>N</i>
Mathematical content	Number and operations (NO)	8
	Algebra (AL)	5
	Geometry and measurement (GM)	5
	Statistics and probability (SP)	7
Problem-solving skills	Interpretation (I)	7
	Strategic knowledge and reasoning (S)	9
	Modeling (M)	8
	Communication (C)	1
Total		25

13.4.3 Analysis

All planned analyses were based on raw scores. To address the first research question regarding the test's psychometric quality, we examined item difficulty, item discrimination, differential item functioning (DIF), reliability, and dimensionality. Item difficulty was calculated as the proportion correct (i.e., the *p*-value) in order to evaluate whether the test was at the appropriate difficulty level for the target population. To evaluate the extent to which the test was able to differentiate between high- and low-performing test takers, item discrimination was evaluated using item-total point-biserial correlations (i.e., uncorrected r_{pbis}).

Furthermore, we conducted reliability analyses both at the test-taker level and at the institutional level to evaluate whether scores were reliable for reporting purposes for individual and institutional use. We used Cronbach's alpha for estimating individual-level reliability. Institutional-level reliability was calculated using a split-sample approach illustrated in Klein et al. (2007). This procedure³ involves randomly splitting the students in each school into Sample A and Sample B, computing mean scores for both samples at each school, and correlating the Sample A and Sample B means across all the schools. A Spearman-Brown correction was used to adjust for the use of half-size samples. In our analyses, the mean of 30 random

³Since there are no clear guidelines for the minimum sample size needed for estimating group-level reliability using this approach, we decided not to exclude schools due to small sample sizes.

splits was computed in order to obtain a stable estimate of the expected value of school-level reliability.

Regarding the internal structure of the test, two sets of analyses were conducted. Observed and disattenuated correlations were reported among the four content areas as well as among the three skills both at the individual and institutional levels. Correlations at the institutional level were calculated using the mean score from each university. Disattenuated correlations are observed correlations adjusted for measurement error. In addition, a confirmatory factor analysis (CFA) approach was taken to examine the latent structure that underlies the relationships among the test items. Three competing models were tested, a unidimensional model, a correlated four-factor content model, and a correlated three-factor skill model. In the unidimensional model, all items load on a single latent factor. This model hypothesizes that the test performance can be accounted for by a single ability factor. In the correlated four-factor model, items within each content area load on their respective content factors, and the four-content factors are correlated with one another. This model hypothesizes that there are four distinct, and yet correlated, content factors. In the correlated three-factor model, items pertaining to each skill area load on their respective skill factors, and the three skill factors are correlated with one another. According to this model, the three skill factors can be statistically differentiated. Latent analyses were based on item-level raw scores using Mplus version 6.1 (Muthén and Muthén 2010). Since all items in the test were dichotomously scored, we used the WLSMV estimator, a robust diagonally weighted least squares (DWLS) estimator provided by Mplus to adjust the parameter estimates, standard errors, and fit indices for the categorical nature of the data as suggested by Finney and DiStefano (2013). Two DWLS-based global fit indices were used for evaluating model fit: (a) comparative fit index (CFI) and (b) root mean square error of approximation (RMSEA). Finney and DiStefano (2013) suggested that guidelines similar to those used for maximum likelihood-based fit indices can apply to DWLS-based indices. Following their suggestions, a CFI value larger than 0.94 and a RMSEA value smaller than 0.06 indicate good model-data fit. Individual parameter estimates were also examined for appropriateness and significance. A latent factor correlation of 0.90 was used to screen out models with extreme factor dependency as this criterion was used in validation studies for other large-scale standardized assessments (e.g., Sawaki et al. 2009).

To address the second research question, that is, to examine the relation between the test and external variables, we identified three types of construct-relevant variables: self-rated QL skills, test-taker perceptions of the testing experience (i.e., perceived test difficulty and testing time), and prior academic success indicators.

Self-rated QL skills were reported on a four-point Likert scale. Test takers were also asked to report perceived test difficulty and whether they had enough time to finish the test, both of which were reported on a three-point Likert scale. Separate one-way ANOVAs were applied to examine the performance differences by self-rated skills, perceived test difficulty, and perceived testing time. Following the finding of a significant main effect, we then conducted post hoc pair-wise comparisons to examine which pairings contributed to the overall statistical difference. The Bonferroni procedure was used for the follow-up analysis to control for the family-wise Type I error due to multiple comparisons.

HEIghTen QL performance was examined in relation to three academic success indicators: (a) university elite status, (b) high school selectiveness, and (c) university entrance exam performance. Separate *t*-tests were used to examine performance differences between those from elite schools and those from non-elite schools, and between those who attended advanced high schools and those who attended regular high schools. Test takers reported their scores on the Russian college entrance exam, the Unified State Exam (USE), for the following four subject areas: mathematics, Russian language, physics, and informatics. Performance was reported on a 100-point scale for each exam. Scores were not comparable across years as the tests from different years are not linked. We performed two kinds of analyses to determine the relationships between USE and *HEIghTen*. We examined the Pearson correlations between USE test scores and the *HEIghTen* QL score by grade. We also used regression analysis to examine the extent to which *HEIghTen* performance could be predicted by USE results. A multiple regression model was tested to predict the *HEIghTen* QL score using USE scores after controlling for grade in a hierarchical fashion. In this model, grade was entered first, followed by the USE scores entered all at once.

Effect size was also reported. Using Cohen’s (1988) guidelines, the criteria for a small, medium, and large *d* are 0.2, 0.5, and 0.8, respectively, and the criteria for a small, medium, and large eta squared (η^2) are 0.01, 0.06, and 0.14, respectively.

13.5 Results

13.5.1 Item Difficulty and Discrimination

The total score ranged from 0 to 25. The mean and standard deviations of the total score were 15.55 and 4.55, respectively. Table 13.3 shows the mean and range of item difficulty and discrimination for the total test, as well as by content and by skill.

Table 13.3 Item difficulty and discrimination

	Difficulty		Discrimination	
	Mean	Range	Mean	Range
Total	0.62	0.16–0.92	0.42	0.19–0.57
Content				
Number and operations	0.71	0.41–0.89	0.42	0.33–0.57
Algebra	0.62	0.34–0.81	0.48	0.45–0.52
Geometry and measurement	0.71	0.30–0.92	0.40	0.32–0.47
Statistics and probability	0.45	0.16–0.73	0.39	0.19–0.57
Skill				
Communication and interpretation	0.70	0.41–0.92	0.42	0.30–0.57
Strategic knowledge and reasoning	0.58	0.16–0.91	0.41	0.19–0.57
Modeling	0.60	0.16–0.81	0.43	0.33–0.52

Note: item difficulty = proportion correct (i.e., *p*-value). Item discrimination = item-total point-biserial correlation

The p -values of the 25 test items ranged from 0.16 to 0.92. An item difficulty range of 0.30 to 0.80 is typically aimed for by existing SLO assessments, such as the Collegiate Learning Assessment + (CLA+) (Council for Aid to Education 2015) and the Collegiate Assessment of Academic Proficiency (CAAP) (ACT 2012). In our data, eight items had an item difficulty larger than 0.80 and three had an item difficulty smaller than 0.30. These items appeared to be either too difficult or too easy for the group. The average item difficulty at the test level was 0.62, which was within the typical difficulty range for existing SLO assessments. This indicated that the test as a whole was at the appropriate difficulty level for the sample. Furthermore, analysis by content and by skill showed that the average difficulties at the subscale level were all within the expected range, suggesting that none of the subscales appeared to be extremely difficult or easy for the sample.

Regarding item discrimination, the average item discrimination at the total test level was 0.42, with point-biserial correlations ranging from 0.19 to 0.57. These are comparable to existing SLOs. For example, a point-biserial correlation of at least 0.10 was used for selecting operational items in the CLA+ item bank (Council for Aid to Education 2015).

13.5.2 Reliability

Total and sub-score reliability estimates calculated at the individual and institutional levels are reported in Table 13.4.

Table 13.4 Individual- and institutional-level reliability

	Institutional-level reliability (CI lower-CI upper)	Individual-level reliability
Total	0.86 (0.79–0.91)	0.81
Content area		
Number and operations	0.73 (0.56–0.85)	0.58
Algebra	0.78 (0.69–0.88)	0.54
Geometry and measurement	0.78 (0.65–0.88)	0.41
Statistics and probability	0.84 (0.75–0.91)	0.49
Skill		
Communication and interpretation	0.78 (0.64–0.89)	0.55
Strategic knowledge and reasoning	0.83 (0.73–0.91)	0.58
Modeling	0.80 (0.67–0.90)	0.58

As the *HEIghten* assessment suite is designed primarily to be used at the group level, we report institutional-level total score and subscale reliability estimates. At the institutional level, the total score reliability was 0.86. In addition, all institution-level subscale reliabilities were above 0.70. These estimates were comparable to group-level estimates reported for existing SLO assessments (e.g., CLA+, CAAP; Klein et al. 2009).

We further explored whether scores were reliable at the individual level. We found that the total score reliability was 0.81, which was comparable to test-level reliability estimates reported for existing SLO assessments. For example, CLA+ was reported to have a reliability of 0.81 for the total score (Council for Aid to Education 2015). The CAAP total score reliability estimates ranged from 0.84 to 0.92 across test forms (ACT 2012). For the ETS[®] Proficiency Profile (EPP) test, the total score reliability estimate was 0.91 for the Standard form and 0.77 for the Abbreviated form (ETS 2010). Across the subscales, the individual reliability estimates ranged from 0.41 to 0.58.⁴

13.5.3 Relations Among the Sub-scores

The observed and disattenuated correlations across the sub-scores at the individual and institution levels are reported in Tables 13.5 and 13.6. After measurement error being adjusted, the disattenuated correlations among the four content areas and among the three skill areas were all very high, indicating strong associations among the sub-constructs.

Table 13.5 Individual-level observed and disattenuated correlations across content and skill areas

Content	NO with AL	NO with GM	NO with SP	AL with GM	AL with SP	GM with SP
Observed	0.57	0.50	0.53	0.50	0.54	0.46
Disattenuated	1.00 ^a	1.00 ^a	0.99	1.00 ^a	1.00 ^a	1.00 ^a
Skill	CI with M	CI with S	M and S			
Observed	0.58	0.61	0.64			
Disattenuated	1.00 ^a	1.00 ^a	1.00 ^a			

Note: NO number and operations, AL algebra, GM geometry and measurement, SP statistics and probability, CI communication/interpretation, S strategic knowledge and reasoning, M modeling
^aValues greater than 1.00 are reported as 1.00

⁴Relatively low reliabilities at the subscale level were expected and should not be a concern because, by design, *HEIghten* scores are not reported to individual students.

Table 13.6 Institution-level observed and disattenuated correlations across content and skill areas

Content	NO with AL	NO with GM	NO with SP	AL with GM	AL with SP	GM with SP
Observed	0.93	0.80	0.86	0.91	0.89	0.83
Disattenuated	1.00 ^a	0.98	0.98	1.00 ^a	1.00 ^a	1.00 ^a
Skills	CI with M	CI with S	M and S			
Observed	0.90	0.89	0.90			
Disattenuated	1.00 ^a	0.98	1.00 ^a			

Note: NO number and operations, AL algebra, GM geometry and measurement, SP statistics and probability, CI communication/interpretation, S strategic knowledge and reasoning, M modeling
^aValues greater than 1.00 are reported as 1.00

Results from the CFA analysis are reported in Table 13.7. All three hypothesized models converged. The selected global fit indices for the three models were all satisfactory, indicating good model-data fit. However, very high latent factor correlations were found in the two multi-factor models, suggesting that the sub-constructs could not be statistically differentiated. As high correlations among the subscales were also found in the US pilot study (Roohr et al. 2017), our finding further confirmed the unidimensional nature of the assessment.

Table 13.7 Results of confirmatory factor analysis

Model	Chi-square	df	CFI	TLI	RMSEA
Unidimensional	519.18	275	0.96	0.96	0.03
Content four factor	517.50	269	0.96	0.96	0.03
Skill three factor	512.29	272	0.96	0.96	0.03
Factor correlations in the four-factor content model					
	No	AL	GM	SP	
NO	1.00				
AL	1.00 ^a	1.00			
GM	1.00	1.00	1.00		
SP	0.98	1.00 ^a	0.98	1.00	
Factor correlations in the three-factor skill model					
	CI	M	S		
CI	1.00				
M	1.00 ^a	1.00			
S	1.00 ^a	1.00 ^a	1.00		

Note: NO number and operations, AL algebra, GM geometry and measurement, SP statistics and probability, CI communication/interpretation, S strategic knowledge and reasoning, M modeling
^aValues greater than 1.00 are reported as 1.00

13.5.4 Relations with Self-Rated QL Skills

The test takers were asked to rate their QL skills on a four-point Likert scale, excellent, good, average, and poor. We considered their self-rated QL skills as an external variable that was relevant to the construct. We therefore examined the relations between their *HEIghten* QL performance and self-rated QL skills.

Results reported in Table 13.8 showed that test takers' self-rated QL skills corresponded to their actual test performance. Test performances across the four rating groups were significantly different, with a close-to-large effect size, $F_{(3, 1165)} = 53.69$, $p < 0.001$, $\eta^2 = 0.121$. As shown in Table 13.8, all follow-up pair-wise comparisons using the Bonferroni procedure were significant ($p < 0.05$). In addition, those who rated themselves excellent or good ($M = 16.89$, $SD = 4.32$) outperformed those who rated themselves average or poor ($M = 13.93$, $SD = 4.26$). This difference was significant with a medium-to-large effect size, $t_{(1167)} = 11.74$, $p < 0.001$, $d = 0.69$.

Table 13.8 Self-rated QL skills and test performance

Self-rated QL skills	N	%	Mean	SD	Cohen's <i>d</i>			
					1	2	3	4
Excellent (1)	88	7.50	18.07	4.19	–			
Good (2)	560	47.60	16.71	4.32	0.32*	–		
Average (3)	474	40.30	14.14	4.25	0.93***	0.60***	–	
Poor (4)	47	4.00	11.77	3.78	1.56***	1.15***	0.56**	–

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

13.5.5 Relations with Perceived Test Difficulty and Testing Time

Test takers were also asked to report perceived test difficulty and whether they had enough time to finish the test, both of which were considered to be construct relevant.

Results reported in Table 13.9 showed that the perceived test difficulty corresponded to the actual test performance. The performances across those who rated the test to be “too easy” (Group 1), “at the right level” (Group 2), and “too difficult” (Group 3) were significantly different with a medium effect size, $F_{(2, 1166)} = 33.27$, $p < 0.001$, $\eta^2 = 0.054$. Results from the follow-up pair-wise comparisons showed that Group 1 ($M = 17.00$, $SD = 4.46$) significantly ($p < 0.001$) outperformed Group 2 ($M = 14.99$, $SD = 4.34$) with an effect size of 0.46 and Group 3 ($M = 12.89$, $SD = 5.43$) with an effect size of 0.90. Also significantly, but to a lesser degree ($p < 0.05$), Group 2 outperformed Group 3 with an effect size of 0.48.

Table 13.9 Perceived test difficulty, testing time, and test performance

	<i>N</i>	%	Mean	SD	Cohen's <i>d</i>		
					1	2	3
Test difficulty							
Too easy (1)	378	32.14	17.00	4.46	–		
At the right level (2)	753	64.03	14.99	4.34	0.46***	–	
Too difficult (3)	38	3.23	12.89	5.43	0.90***	0.48*	–
Testing time							
More than enough (1)	285	24.23	15.96	4.72	–		
Enough time (2)	593	50.43	15.79	4.56	0.04	–	
Not enough time (3)	291	24.74	14.74	4.20	0.27**	0.24**	–

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The relationship between test performance and perceived testing time was also in the expected direction. As shown in Table 13.9, the performances across the three groups who responded “more than enough time” (Group 1), “enough time” (Group 2), and “not enough time” were significantly different with a small effect size, $F_{(2, 1166)} = 6.74$, $p < 0.01$, $\eta^2 = 0.011$. Results from the follow-up pair-wise comparisons showed that Group 3 ($M = 14.74$, $SD = 4.20$) performed significantly worse than Group 1 ($M = 15.96$, $SD = 4.72$) with an effect size of 0.27 and Group 2 ($M = 15.79$, $SD = 4.56$) with an effect size of 0.24 significantly ($p < 0.01$). No significant difference was found between groups 1 and 2 ($p > 0.05$).

13.5.6 Relations with Prior Academic Success

One indicator of prior academic success was university elite status. If the test functions well, we would expect that students from elite universities would outperform those from non-elite schools. The results confirmed our expectation. Students from elite universities ($M = 16.94$, $SD = 4.54$) performed significantly better than those from non-elite universities ($M = 15.20$, $SD = 4.49$) with a small-to-medium effect size, $t_{(1174)} = 5.29$, $p < 0.001$, $d = 0.38$.

Another indicator was high school type. Some reported having attended advanced schools⁵ while the others reported having attended regular schools. The difference between the groups was in the expected direction. Students from advanced high schools ($M = 16.58$, $SD = 4.45$) outperformed those from regular high schools ($M = 15.10$, $SD = 4.50$); $t_{(1167)} = 5.28$, $p < 0.001$, $d = 0.33$.

The third indicator we used was performance on the USE tests. The students reported their scores for four of the USE tests, mathematics, Russian language,

⁵In the Russian educational system, advanced high schools offer educational programs of higher level than regular high schools.

physics, and informatics. We examined the relations between *HEIghten* QL performance and scores on the USE tests.

Pearson correlations were calculated for freshmen and juniors separately because, as mentioned earlier, USE tests were not comparable from year to year. The results are presented in Table 13.10. All correlations were significant. In addition, QL scores had stronger associations with mathematics and informatics, than with Russian language and physics. The results were aligned with our expectation as we expected that *HEIghten* scores would not only relate positively to USE tests but also relate more strongly with tests that measure constructs similar to the *HEIghten* QL assessment (e.g., mathematics) than those that measure different constructs (e.g., Russian language).

Table 13.10 Correlations with USE tests by grade

	Year	<i>N</i>	Pearson’s correlation
Math (<i>N</i> = 943)	Freshmen	517	0.507***
	Juniors	426	0.387***
Russian (<i>N</i> = 943)	Freshmen	516	0.423***
	Juniors	427	0.294***
Physics (<i>N</i> = 712)	Freshmen	379	0.400***
	Juniors	333	0.285***
Informatics (<i>N</i> = 471)	Freshmen	273	0.531***
	Juniors	198	0.447***

Note: ****p* < 0.001

We also examined the extent to which *HEIghten* QL performance could be predicted by the USE tests. A multiple regression analysis was conducted in which model testing was carried out in two sequential steps. Grade was entered first (Model 1), followed by adding the USE scores to the model (Model 2). As shown in Table 13.11, when the USE scores were entered in Model 2, the variance explained increased to 28.7% with a significant *F* change. The inclusion of the USE scores as predictors accounted for an additional 28.4% of the total variance of the test performance.

Table 13.11 Model summary of multiple regression

Model	<i>R</i>	<i>R</i> square	Std. error of the estimate	<i>R</i> square change	<i>F</i> change	<i>df</i> 1	<i>df</i> 2	Sig. <i>F</i> change
1	0.051	0.003	4.556	0.003	0.654	1	252	0.419
2	0.536	0.287	3.883	0.284	24.716	4	248	0.000

Table 13.12 shows the estimated model parameters. After controlling for grade, all USE tests except for physics contributed significantly to model prediction, with mathematics and informatics being the two strongest predictors.

Table 13.12 Regression coefficients from multiple regression

		Standardized coefficients	<i>t</i> -test	<i>p</i> -value
Step 1				
Intercept	15.108		24.614	0.000
Grade	0.233	0.051	0.809	0.419
Step 2				
Intercept	1.286		0.747	0.456
Grade	-0.569	-0.124	-2.088	0.038*
Math	0.067	0.216	2.923	0.004**
Russian	0.065	0.157	2.534	0.012*
Physics	0.002	0.005	0.075	0.940
Informatics	0.097	0.309	4.093	0.000***

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

13.6 Discussions and Conclusion

The purpose of the study was to investigate validity evidence regarding the use of translated and adapted *HEIghten* QL assessment in Russia. Including only EE and CS majors in our study sample may limit the generalization of the study results to Russian college students majoring in other fields. With this caveat in mind, we found convincing evidence in support of the use of the Russian *HEIghten* QL assessment.

Generally speaking, the assessment had appropriate psychometric quality for the target population as a measure of QL. The test was found to be at the appropriate difficulty level and be able to differentiate test takers well. The total score reliability was acceptable at both individual and institutional levels, providing initial support for reporting the total score for both individual and institutional uses. In addition, the sub-score reliabilities at the institutional level were also satisfactory, suggesting that sub-scores can be reported to help identify strengths and weaknesses of schools. We also found that the assessment was practically unidimensional in nature. On the one hand, this finding was consistent with the results from the *HEIghten* QL pilot study conducted with higher education institutions in the US. On the other hand, this finding also suggested that sub-scores may provide little distinct information that is not already included in the total score, a common concern regarding the use of diagnostic sub-scores (e.g., Haberman 2008). To further investigate the value of sub-scores in addition to the total score, we need to examine whether the same internal structure holds across different institutional types and/or diverse learner groups to understand whether sub-scores can potentially be used by some sub-groups of the test-taking populations for diagnostic purposes.

In addition to validation evidence based on the test's internal structure, the test was also found to have appropriate relationships with different types of construct-relevant external variables. This outcome not only contributed evidence in support of the use of the test but also provided insight into the potential factors associated

with QL achievement in Russia. In particular, two variables that were characteristic of the Russian context, university elite status and high school type, were both strongly associated with performance measured by the *HEIghten* QL assessment. One caveat in this analysis was that we were not able to control for prior achievement when examining performance differences across university types or high school types. Nevertheless, the finding can facilitate the identification of relevant contextual variables for learning. For example, Trigwell and Prosser (1991) and Lizzio et al. (2002) found that both student perceptions and evaluations of learning environment and their approaches to study related to learning outcomes. Multiple predictors of academic success in higher education, including intellectual ability, learning style, personality, and achievement motivation, were examined in Busato et al. (2000). Further analysis is needed to explain what characteristics (e.g., learning environment, teaching approach) of the different types of universities and high schools could have contributed to the observed performance differences. Research that explores the relationships between characteristics at the country, institution, and person level would contribute to an understanding of how to improve the quality of learning across diverse international contexts.

Two study limitations are worth noting. First, in this study we focused only on two types of validity evidence, namely, the test's internal structure and the test's relations with external variables. To develop a convincing and coherent validity argument for the use of this assessment in Russia, future studies should explore the other sources of validity evidence that are deemed essential by the *Standards* (AREA, APA, and NCME 2014). For example, evidence based on test content can be investigated by examining the alignment between standards and practices for math education in Russia and the *HEIghten* QL assessment framework. Evidence based on response process should also be investigated because the cognitive processes engaged in by test takers when responding to test items could differ between English speakers, for whom the assessment was originally developed, and Russian speakers because of differences in their educational, social, and cultural backgrounds. Furthermore, the consequences of using translated and adapted tests in an international context, both intended and unintended, should be evaluated. The second study limitation relates to the unique demand that learning outcome assessments are designed to fulfill, that is, to be able to assess change as a function of learning. Although we found convincing evidence to support the use of the assessment as a measure of QL in Russia, we were not able to investigate its utility as a measure of gains in QL as the result of learning. To argue that this measure is capable of measuring change due to learning would require longitudinal test performance data and associated information on learning, to which we did not have access at the time of this validation study. As a critical area for future research, we suggest investigating validity evidence that links changes in test performance to relevant learning experiences to verify the value of the assessment of measuring student learning outcomes.

Acknowledgments Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

References

- American College Testing. (2012). *ACT CAAP technical handbook 2011–2012*. Iowa City: CAAP Program Management. Retrieved from <http://www.act.org/content/dam/act/unsecured/documents/CAAP-TechnicalHandbook.pdf>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bennell, P., & Pierce, T. (2003). The internationalisation of tertiary education: Exporting education to developing and transitional economies. *International Journal of Educational Development*, 23, 215–232.
- 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, 1057–1068.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: L. Erlbaum Associates.
- Council for Aid to Education. (2015). *CLA+ technical FAQs*. New York: Author. Retrieved from http://cae.org/images/uploads/pdf/CLA_Plus_Technical_FAQs.pdf
- Educational Testing Service. (2010). *ETS proficiency profile user's guide*. Princeton: Educational Testing Service. Retrieved from https://www.ets.org/s/proficiencyprofile/pdf/Users_Guide.pdf
- Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte: Information Age.
- Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33, 204–229.
- International Test Commission. (2005). *International Test Commission guidelines for translating and adapting tests*. Retrieved from http://www.intestcom.org/files/guideline_test_adaptation.pdf
- Klein, S., Benjamin, R., Shavelson, R. J., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review*, 31, 415–439.
- Klein, S., Liu, O. L., Sconing, J. A., Bolus, R., Bridgeman, B., Kugelmass, H., & Steedle, J. (2009). *Test Validity Study (TVS) Report* (ETS technical report). Supported by the Fund for Improvement of postsecondary education (FIPSE). Retrieved from https://cp-files.s3.amazonaws.com/26/TVSReport_Final.pdf
- Knight, J. (2003). Updated definition of internationalization. *International Higher Education*, 33, 2–3.
- Lizzio, A., Wilson, K., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: Implications for theory and practice. *Studies in Higher Education*, 27, 27–52.
- Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles: Authors.
- Organization for Economic Co-operation and Development. (2008). *Tertiary education for the knowledge society*. Paris: OECD. Retrieved from <https://oecd.org/edu/tertiary/review>.
- Organization for Economic Co-operation and Development. (2012). *PISA 2012 technical report*. Paris: OECD. Retrieved from <https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>.
- Roohr, K. C., Graf, E. A., & Liu, O. L. (2014). *Assessing quantitative literacy in higher education: An overview of existing research and assessments with recommendations for next-generation assessment* (ETS RR-14-22). Princeton: Educational Testing Service.
- Roohr, K. C., Lee, H., Xu, J., Liu, O. L., & Wang, Z. (2017). A preliminary evaluation of the psychometric quality of *HEIghTen*TM quantitative literacy. *Numeracy*, 10(2), 3.
- Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL internet-based test. *Language Testing*, 26, 5–30.
- Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education learning outcomes (AHELO) feasibility study report: Design and implementation* (Vol. 1). Paris: Organization for Economic Co-operation and Development.
- Trigwell, K., & Prosser, M. (1991). Improving the quality of student learning: The influence of learning context and student approaches to learning on learning outcomes. *Higher Education*, 22, 251–266.

Chapter 14

Comparative Study of Student Learning and Experiences of Japanese and South Korean Students



Reiko Yamada

Abstract Currently, gains in learning outcomes of college students also become the major theme for higher education institutions worldwide. This research explores to grasp the association of college experiences with degree of learning through the comparative research for student self-reported survey between Japan and Korea. This study uses a quantitative research design using data obtained from JCSS2012 and KCSS2012 designed for upper division students. The research framework, based on five research questions, is to examine the relationship between learning environment students' experiences and learning outcomes between academic majors. We use the KCSS2012 which consists of a stratified random sample of junior and senior students attending four-year universities in South Korea. We finally use 4902 third-year students of private four-year institutions. JCSS2012 consists of samples of junior and senior students attending four-year universities in Japan. We finally use 2921 of both third- and fourth-year students of four-year institutions. Findings of the study suggest that there is a difference of gains of learning outcomes between Japanese and Korean students. Also, the findings suggest that student and faculty engagement variables appear to play important roles in acquisition of knowledge and skills such as globalized skills, interpersonal skills, and cognitive ability. Finally, the finding delineates while many Japanese students have less confidence in their skills and ability, Korean students relatively have more confidence but they have more negative experiences.

This chapter was originally published as the title of “comparative study of learning and student experiences of Japanese and Korean College Students” in *Research in Higher Education*, No. 47, 2015 pp. 169–184 in Japanese. The paper was revised for this book.

R. Yamada (✉)
Doshisha University, Kyoto, Japan
e-mail: ryamada@mail.doshisha.ac.jp

14.1 Introduction

Currently, quality assurance of higher education institutions and enhancement of global competitiveness have become a major concern worldwide. In such an environment, gains in learning outcomes of college students have also become a major theme for higher education institutions worldwide (e.g., Coates, Chap. 1 in this volume). In fact, in recent decades, many institutions across nations have been forced to embed learning outcomes into their curriculum. Japanese higher education institutions have endeavored to develop first-year experience programs and to link faculty development with learning outcomes. In such an environment, universities have frequently used direct assessments represented by portfolio and rubric (Matsushita 2014) as well as self-reported student surveys (Yamada 2012).

These trends are commonly observed in many higher education institutions around the world (e.g., Shavelson et al., Chap. 10 in this volume). Thus, research and practices of measurement or assessment of learning outcomes are being developed worldwide. This also holds true for Korea which is one focus of this chapter. Many South Korean universities have introduced new programs such as the university-college program and first-year experience programs, which are expected to improve learning outcomes. Furthermore, much research is being carried out to develop methods for measuring learning outcomes (cf. Hahn et al., Chap. 8 in this volume). The Korea Research Institute for Vocational Education and Training developed an assessment tool for assessing core skills, which is being used by many higher education institutions (Rhee 2013). In particular, there is much research focused on clarifying the factors that determine the acquisition of generic skills through university education (Choi and Rhee 2009; Choi et al. 2009). Since these self-reported surveys focused exclusively on Korean students, it is difficult to conduct comparative international studies using these data. As such, we revised the JCSS (Japanese College Student Survey) and, then, created a Korean version of the survey (KCSS, Korean College Student Survey) to enable comparative study between Japan and South Korea.

This chapter explores the relationship between students' experiences and learning outcomes through a comparative study of self-reports by Japanese and South Korean university students.

14.2 Literature Review for the Related Study

In what areas have Japanese students' acquisition of abilities and skills improved? This question has become the target of investigation in the context of examining learning outcomes as part of efforts aimed at quality insurance of university education. Recent research on this issue includes a few studies using data from a national survey conducted by the University of Tokyo (Morozumi 2009; Tanimura 2010). This research has revealed that students' active learning and study time outside of

class contribute to students' acquisition of abilities and skills and that the characteristics of classroom instruction (e.g., problem-solving-based or participatory classes) influence the degree to which active learning and study time outside of class are promoted.

Studies have compared the acquisition of abilities and skills in different areas of study. Furuta (2010) observed characteristics in the acquisition of different types of abilities and skills classified in terms of two dimensions—liberal arts-type or scientific-type knowledge and skills—and found that students in liberal art departments tended to give themselves high marks in self-evaluations for acquisition of general knowledge and skills considered to be useful for work. Yamada and others have mainly focused on differences in individual abilities and skills related to learning outcomes acquired by students in three areas-of-study (liberal arts, social sciences, and STEM) as a part of their university life from the three perspectives of educational factors on the university side, factors related to students' efforts, and factors related to students' activities. Although the factors influencing the acquisition of a certain type of ability or skill are not uniform across areas-of-study, these studies have found that proactive commitment to classes and study or reading time outside of class have a positive effect on abilities and skills acquisition while, surprisingly, class attendance does not. Meanwhile, the studies have also highlighted the struggle and long study times of STEM (science, technology, engineering, and mathematics) students both in and outside of class (Sugitani et al. 2013; Yamada 2014a, b).

Much attention has also been paid to identifying areas in which students in other countries have shown improved acquisition of abilities and skills in the context of enhancing quality assurance of university education. In recent years, the focus of research on college impact as related to learning outcomes in the United States has been expanding from just curriculum and university-side environmental factors such as the area-of-study to also include perspectives such as students' engagement and experience.

Pike and Kuh (2005) classified student engagement based on five dimensions: active and collaborative learning, interaction with faculty members, level of academic challenge, enriching educational experiences, and supportive campus environment. Pascarella and Terenzini (2005) analyzed the relationship between these five dimensions and learning outcomes and found that students' active learning is the variable with the greatest impact on learning outcome.

Similarly, Kim and Rhee (2003) demonstrated that, in Korea, greater engagement by a student in his or her instructor's research and greater frequency of interaction with instructors results in higher learning outcomes. Whereas previous research in various countries has demonstrated, to a certain degree, that educational factors on the university side and factors related to students' satisfaction influence students' acquisition of abilities and skills, empirical comparative research on the relationship between students' activity and experience, which includes students' study, and abilities and skills acquisition is just starting. It is expected that more research will be conducted in this area.

Researchers have studied not only the acquisition of abilities and skills but also the nature of university life experienced by different types of students. Mizokami

(2009) performed cluster analysis and identified different student types based on how they perceive study both in and outside the classroom in the context of their overall university experience and investigated how student type was related to differences in student outcome and development.

14.3 Research Objectives

To overcome the shortcoming that the majority of research on students' learning outcomes and experience has been carried out in just single country, since 2010, I have conducted collaborative research on the learning activities and experiences of university students in Japan and Korea. The justification for conducting common surveys and research in the two countries has to do with similarities between the two countries, which include the fact that (1) both countries have undertaken efforts to reform university education in recent years and (2) private universities have played an important role in the shift toward universal advancement to university in both countries. We understand that cross-national assessments are complex in terms of design, measurement of scores, and interpretation of analysis. I and my research partner, engaged in this cross-national comparison, had several meetings for designing the common survey and selecting items which are applicable to students in both countries.

Rhee (2013) conducted surveys based on the same questionnaire (JCSS2010 and KCSS2012) in both Japan and Korea and identified common and disparate factors influencing the acquisition of generic skills by Japanese and South Korean students. In both countries, instruction methods that encourage proactive engagement were found to positively influence the acquisition of generic skills by students classified as active learners. With regard to differences between countries, whereas student-instructor interaction was found to influence skill acquisition in Japan, no such influence was observed in South Korea. That said, in addition to proceeding with data analysis without controlling for differences in the structure of Japanese and Korean universities or students' year in school, the study does not address potential differences in experience that depend on student type. Therefore, in this study, analyzing data from the survey of Japanese and Korean university students conducted in 2012, I clarify the commonalities and differences between Japanese and Korean students in the 3rd and 4th year of study at private universities, while controlling for area-of-study.

The specific objectives of the study are to investigate the following: (1) What are the abilities and skills that the students themselves believe they have acquired (based on self-evaluations)? (2) How do students' experiences and acquired abilities and skills differ by area-of-study? (3) What is the relationship between instruction methods that encourage active learning and students' experience or learning outcomes? And (4) what factors determine students' achievement of learning outcomes? While these objectives are an extension of previous research in this field, the comparison between Japanese and Korean students is new. As I explained before, items of self-reported students survey were carefully discussed and selected between two coun-

tries, each item was carefully translated into local languages, and these translated items were again checked in both countries. Thus, these self-reported student surveys can reflect culture of each country and be compared and analyzed in common framework.

Next, following the approach by Mizokami, I classify the Japanese and Korean students into types based on self-reporting on study, daily life, and self-perception and investigate how students' experiences are influenced by student type.

14.4 Research Methods

14.4.1 *Sample Data and Analysis*

For the Japanese data, I used the JCSS2012 survey whose sample comprised 5786 students in 57 departments at 26 national, public, and private universities. The students included underclassmen in their first and second years of study as well as upperclassmen in their third and fourth years of study in various areas-of-study including liberal arts, social sciences, STEM (science, technology, engineering, and mathematics), health and medicine, and education. For the Korean data, I used the KCSS2012 survey whose sample comprised 6666 third-year students at 51 national and private universities in various areas-of-study including liberal arts, social sciences, STEM, health and medicine, and education.

In this study, I focused on private universities both in Japan and South Korea by using data for third- and fourth-year students attending private universities in both countries in liberal arts, social science, STEM, health and medicine, and education fields. The sample used for analysis comprised 2921 Japanese students in their third or fourth years of study (liberal arts, 287; social sciences, 1150; STEM, 410; health and medicine, 298; education, 438; other fields, 338) and 4902 Korean students in their third year of study (liberal arts, 1060; social sciences, 1419; STEM, 1384; health and medicine, 142; education, 254; other fields, 643). I was unable to control the sample size in each area-of-study, which is a limitation of conducting an international comparative study. Also, while the samples and the size of Korean students were well controlled to reflect proportion of public and private universities in Korea, Japanese samples consist of students' samples of those universities which voluntarily expressed to participate in the survey. This difference made a limitation of analysis of complex cross-national comparison.

Contextualizing this study within the framework of the Input-Environment-Output (I-E-O) model used in college impact studies, I consider high school grade point average (GPA) (performance) and method of entrance to university as attributes as input (I) factors and the students' direct experience or experience mediated by their instructors as environment (E) factors and focus on understanding the relationship between these factors and outcomes in terms of the students' affective and cognitive development. I focused my attention particularly on the relationship between high school GPA and students' experience as well as the quantity and quality of

learning, which is especially interesting given the sample of Japanese and Korean students at private universities. In addition, with respect to environment (E) factors, I was particularly interested in the relationship between students' direct experience or experience mediated by instructors and the quantity and quality of learning, if such relationships vary by area-of-study, and how differences in the quantity and quality of learning stemming from area-of-study or students' experience influence student outcomes in terms of affective and cognitive development. The variables used for analysis for Japanese and Korean private university students included high school GPA, area-of-study, educational content, method of instruction, faculty engagement, students' efforts inside and outside of class, in-class study time (class attendance), study time outside of class, extracurricular activities, and students' self-evaluation of learning outcomes. For typing of students, I conducted factor analysis of students' self-evaluations regarding their abilities and skills compared to other students, etc., and performed cluster analysis using the scores for each factor.

14.4.2 Explanation of Dependent Variables

The questionnaire asks students about their acquisition of relatively generic skills since entering university. Results for a subset of the 20 questions on learning outcomes broken down by area-of-study are presented in Table 14.1. It can be seen that, in general, Korean students gave themselves higher marks in self-evaluations for the acquisition of knowledge and skills. In particular, the proportion of students responding that their knowledge and skills "have improved greatly" was substantially higher for Korean students than for Japanese students. The knowledge and skills that students in both countries ranked highest in terms of improvement were those directly related to university education or the university curriculum and included "discipline-specific knowledge," "general knowledge," and "analytical and problem-solving skills." Students' self-evaluations were also relatively high for skills that might have been acquired through club or other extracurricular activities and experiences outside the university such as the "ability to develop interpersonal relationships" and "ability to collaborate with others." Conversely, the skills that students in both countries ranked lowest in terms of improvement were those that could be characterized as issues of modern society and included the "understanding of global issues," "understanding of local issues," and "understanding of national issues."

To summarize the characteristics of each country, although not shown in the table, the skill with the lowest rating among Japanese students was "foreign language skills," which could be considered a basic academic skill. In contrast, the skill with the lowest rating among Korean students was "understanding of local issues." A substantial difference was observed between Japanese and Korean students regarding "foreign language skills," with the self-evaluation of Korean students being substantially higher than that of Japanese students. It is frequently pointed out that the job market in South Korea for university graduates is extremely competitive and that the students put a lot of effort into acquiring foreign languages with the

Table 14.1 Self-evaluation of Japanese and Korean students' acquisition of abilities and skills after entry into university

	Japan						Korea					
	Lower	N	Middle	N	Upper	N	Lower	N	Middle	N	Upper	N
<i>High school GPA</i>												
General knowledge	65.9	539	71.7	1678	70.9	283	78.6	696	81.8	2622	87.5	684
Analytical and problem-solving skills	61.6	504	68.2	1081	70.7	282	76.0	673	77.4	2481	79.5	622
Discipline-specific knowledge	75.1	614	79.9	1265	81.7	324	91.0	805	92.8	2976	94.1	736
Critical thinking skills	55.4	451	55.5	877	59.8	238	72.7	643	73.3	2349	79.4	621
<i>College GPA</i>												
General knowledge	63.2	573	74.8	1192	69.5	155	77.2	1069	84.0	2756	86.3	189
Analytical and problem-solving skills	59.6	540	71.1	1132	71.5	158	72.3	1000	79.5	2611	81.3	178
Discipline-specific knowledge	70.5	639	83.5	1326	82.0	183	89.1	1232	94.0	3088	94.5	208
Critical thinking skills	50.8	458	59.5	944	62.6	139	68.6	950	76.3	2508	76.7	168

(Total % of greatly increase and increase)

$p < 0.001$

goal of expanding their career opportunities beyond the limited domestic job market to a more global market.¹ As such, many students study foreign languages outside the university. Thus, it is necessary to keep in mind that this result may not be due exclusively to university curricula.

In general, the students perceived that they had acquired knowledge and skills specific to their chosen areas-of-study. For liberal arts students, the acquired skills were related to the humanities and included “general knowledge,” “knowledge of people from different races/cultures,” “writing ability,” and “foreign language skills.” For students in the social sciences, the acquired skills were related to social issues and included “understanding of national issues” and “understanding of local issues.” For STEM students, the acquired skills were related to analysis and quantitative treatment and included “mathematical ability,” “analytical and problem-solving skills,” and “IT skills.” Although differences in knowledge acquisition were observed between Japanese students in different areas-of-study, very few differences were observed among Korean students in different areas-of-study (Fig. 14.1).²

Principal component analysis (varimax method) of the 20 abilities and skills that changed after entry into university resulted in the identification of three components (factor loading of 0.400 or greater; cumulative contribution ratio: 61.2%), which I named *global competency* ($\alpha = 0.859$), *interpersonal skills* ($\alpha = 0.821$), and *cognitive ability* ($\alpha = 0.757$)

14.5 Results

14.5.1 Student Background and Acquisition of Abilities and Skills After Enrollment in University

The relationships between the students’ high school GPA (which is an input factor) or students’ current GPA and the variables contributing to the *cognitive ability* component identified by factor analysis are shown in Table 14.1.

A higher proportion of the top group of students in both Japan and South Korea in terms of both high school GPA and university GPA responded that their abilities and skills had improved in all areas. However, for all items, the proportion of students who reported that their abilities and skills had improved was significantly higher among Korean students than Japanese students, marking a clear difference between Japanese and Korean students.

¹View expressed in informal discussions with multiple university instructors during interview surveys carried out in 2012 and 2013 in South Korea.

²The variables contributing to global competency include [understanding of global issues], [ability to work with people from different races/cultures], [knowledge of people from different races/cultures], [foreign language skills], [understanding of national issues], and [understanding of local issues]. The variables contributing to interpersonal skills include [ability to develop interpersonal relationships], [ability to collaborate with others], [leadership skills], and [time management skills]. The variables contributing to cognitive ability include [discipline-specific knowledge], [general knowledge], [analytical and problem-solving skills], and [critical thinking skills].

Japan				Korea		
67		Humanities	knowledge of discipline and major	Humanities		93.1
77.1		Social Sciences		Social Sciences		91.9
74.1		STEM		STEM		92.1
88.6		Medical and Nursing		Medical and Nursing		94.4
82.3		Teacher training	Teacher training	Teacher training		92.1
73.1		Humanities	general knowledge	Humanities		84.3
71.4		Social Sciences		Social Sciences		83.1
61.2		STEM		STEM		80.2
67.2		Medical and Nursing		Medical and Nursing		83.1
73.4		Teacher training	Teacher training	Teacher training		80.2
65.2		Humanities	analytical and problem solving skills	Humanities		75.9
66.7		Social Sciences		Social Sciences		75.7
65.2		STEM		STEM		78.9
69.6		Medical and Nursing		Medical and Nursing		81
67.2		Teacher training	Teacher training	Teacher training		79.6
68.7		Humanities	interpersonal skills	Humanities		74
67.2		Social Sciences		Social Sciences		76.9
52.6		STEM		STEM		78
71.6		Medical and Nursing		Medical and Nursing		81.7
72.7		Teacher training	Teacher training	Teacher training		77.2
68.3		Humanities	knowledge of people from different races/cultures	Humanities		60.6
50.4		Social Sciences		Social Sciences		58.4
35.5		STEM		STEM		41.8
26		Medical and Nursing		Medical and Nursing		40.8
49.9		Teacher training	Teacher training	Teacher training		58.3
54.6		Humanities	Preparedness for career after college	Humanities		51.6
52.4		Social Sciences		Social Sciences		64.7
42		STEM		STEM		64.7
54.4		Medical and Nursing		Medical and Nursing		76.8
65.4		Teacher training	Teacher training	Teacher training		61.4

Fig. 14.1 Self-evaluations of acquired abilities and skills by Japanese and Korean students

14.5.2 Difference in Activity Times and Experience of Japanese and Korean Students Broken Down by Area-of-Study

Previous research using JCSS data has shown that substantial differences exist between areas-of-study in terms of factors such as time spent studying, curriculum, and instruction methods. Given that research up to this point has not examined whether or not there are differences in study time or university experiences between countries depending on area-of-study, I examined how the two factors of country (in this case, Japan and South Korea) and area-of-study are related to amount of learning. Table 14.2 shows the results of two-way ANOVA of country and area-of-study on mean study time, which indicate that the main effects of country and area-of-study are significant. Mean study time for all areas-of-study except for health and medicine was found to be longer for Korean students than for Japanese students. While the results must be interpreted carefully since the interaction between country and area-of-study was found to be significant, it was confirmed that, to a certain degree, differences exist between countries.

Table 14.2 Two-way ANOVA of study time outside class for Japanese and Korean students by area-of-study

Disciplines	Humanities		Social Sciences		STEM		Medical and Nursing		Teacher training		Main effect		Interaction
	Japan	Korea	Japan	Korea	Japan	Korea	Japan	Korea	Japan	Korea	Country	Discipline	
Average	3.2	4.7	3.1	4.4	3.7	4.8	4.4	4.4	3.1	4.7	464.23***	29.33***	16.76***
SD	1.44	1.58	1.52	1.49	1.74	1.7	2.03	1.55	1.52	1.55			

*** $p < 0.001$ average 1 = 0 Hour 2 = 0.5H 3 = 1.5H. 4 = 4H 5 = 8H 6 = 13H 7 = 18H 8 = over 20H

14.5.3 *Relationship Between Active Learning and Learning Outcomes*

In previous research on the relationship between experience with active learning methods and the acquisition of analytical and problem-solving skills, over 70% of students at national, public, and private universities who had experienced “presenting my own thoughts and research in class” reported that their “analytical and problem-solving skills” had improved, whereas less than 30% of students who did not have such experience reported similar improvement. Furthermore, over 70% of students who had experience “presenting my own thoughts and research in class” reported that their “communication skills,” “presentation skills,” and “discipline-specific knowledge” had improved, indicating that active learning methods, to a certain degree, promote students’ proactive learning and acquisition of abilities and skills related to undergraduate curricula (Yamada 2014a, b).

Results of my analysis of the relationship between the level of active learning experience in different areas-of-study and learning outcomes in Japan and South Korea are presented in Table 14.3.³ For Japanese students, the active learning experience of “having discussions with other students in class” had a significant main effect on discipline-specific knowledge and engagement but not *cognitive ability*. No interaction effect was observed for any pair of factors. A high proportion of students who had the opportunity to “have discussions with other students in class” reported improvement of *global competency*, *interpersonal skills*, and *cognitive ability*, with some variation between areas-of-study. A higher proportion of students in liberal arts, social science, and education-related majors reported improvement of *global competency* and *interpersonal skills* than students in STEM or health- and medicine-related majors. The difference between areas-of-study in terms of improvement of *cognitive ability* was small.

For Korean students, active learning experience had a significant main effect on engagement but not on areas-of-study. As in the case of the Japanese students, no interaction effect was observed for any pair of factors. Except for *global competency* for students in health- and medicine-related majors, a higher proportion of Korean students who had the opportunity to “have discussions with other students in class” reported improvement of learning outcomes. The pattern of differences between areas-of-study varied from and the magnitude of differences were smaller than those observed in Japan. That said, the above analysis does not take the inter-correlation of variables into account. In the next section, I investigate the influence of environmental (E) factors and input (I) factors on affective and cognitive development using multiple regressions.

³The table only includes items that were contained in both the Japanese and Korean surveys and for which significant differences were observed.

Table 14.3 Two-way ANOVA of learning outcomes on active learning experience and discipline in Japan and South Korea

	Humanities		Social sciences		STEM		Medical and nursing		Teacher training		Main effect		
	Opportunity	No opportunity	Opportunity	No opportunity	Opportunity	No opportunity	Opportunity	No opportunity	Opportunity	No opportunity	Discipline	Engagement	Interaction
<i>Japan: having discussions with other students in class</i>													
Global competency	18.3	17.1	17.4	16.5	16.6	15.8	16.1	15.3	17.0	16.6	31.376**	50.618***	0.727
Interpersonal skills	3.141	3.192	2.757	2.958	2.590	2.647	2.336	3.098	2.507	2.822	8.496*	43.224**	1.318
Cognitive ability	3.751	3.500	3.210	3.463	2.978	3.484	2.621	3.264	2.895	3.440	1.910	52.092***	0.965
	2.304	2.517	2.176	2.422	2.109	2.648	1.963	2.269	1.999	2.425			
<i>Korea: presenting my own thoughts and research in class</i>													
Global competency	18.0	17.0	18.4	18.0	17.3	16.0	17.0	17.1	17.8	14.3	5.166	7.055*	1.637
Interpersonal skills	2.984	2.887	2.860	3.698	3.048	2.811	2.793	3.603	3.271	1.500	1.923	16.842***	0.619
Cognitive ability	3.366	4.059	3.425	3.623	3.425	3.621	3.409	2.712	3.775	1.258	0.814	16.998**	2.105
	2.004	2.083	1.938	2.646	2.067	2.600	2.042	2.031	1.918	1.500			

Upper, average; lower, SD

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

14.5.4 Predictors of Skill Acquisition

For independent variables, the regression model included method of entrance to university, high school GPA, and gender as input (I) factors, and university GPA, area-of-study, variables related to student experience and faculty engagement, study time outside of class and hours of class attendance as variables related to educational quantity, and student satisfaction as a proxy for educational quality (since no direct measurement exists) as educational environment (E) factors. Table 14.4 presents the results of multiple regression analyses with *global competency*, *interpersonal skills*, and *cognitive ability* as outcome variables.

In terms of common results observed for both Japanese and Korean students, students' satisfaction, representing the students' overall experience at university, resulted in improvement of *global competency*. Furthermore, a higher proportion of students in liberal arts- and social science-related majors reported improvement of *global competency* than students in health- and medicine-related majors. The influence of input (I) factors was found to differ between Korean and Japanese students. Male Korean students had a greater tendency to report improvement of *global competency* than female Korean students, while Japanese students with higher grades had a greater tendency to report improvement of *global competency* than Japanese students with lower grades.

With respect to the influence of factors related to students' experience, whereas "having discussion with other students in class" was found to have a negative influence on *global competency* among Japanese students, the impact was found to be positive among Korean students. It should be noted that, although Table 14.3 indicated that a high proportion of Japanese students who had the opportunity to have discussions with other students in class reported improvement of *global competency*, this result is contradicted—i.e., the influence of active learning experience is found to be negative—when other factors are included in multiple regression analysis.

Furthermore, whereas study time outside of class was found to have a negative influence on *global competency* among Japanese students, the impact was found to be positive among Korean students. The influence of factors related to students' experience was found to differ substantially between Japanese and Korean students. For Japanese students, proactive commitment to reading and participation in study abroad programs had a positive influence on *global competency*. In contrast, for Korean students, attendance of classes in which the students themselves research literature and other materials had a positive influence on *global competency*.

With respect to variables related to faculty engagement, instructors' encouragement to pursue graduate school or professional studies had a positive influence on *global competency* for students in both countries. Although the specific nature of faculty engagement differed between countries, it is evident that faculty engagement had an influence on *global competency*. For Japanese students, the opportunity to talk about class content outside of class had a positive impact on *global competency*, whereas for Korean students, instructors' emotional support had a positive impact on *global competency*.

Table 14.4 Factors predicting learning outcomes: multiple regression analyses^a

Outcome: Global competencies	Model 1		Model 2		Model 3		Model 4	
	Japan	Korea	J	K	J	K	J	K
Attribute	b	b	b	b	b	b	b	b
<i>Gender dummy</i>	-0.027	0.142***	0.004	0.105***	-0.002	0.099***	-0.005	0.091***
Humanities dummy	0.164***	0.063*	0.109***	0.061*	0.11***	0.058*	0.115***	0.072***
Social sciences dummy	0.103***	0.118**	0.114***	0.137***	0.114***	0.138***	0.135***	0.159***
STEM dummy	-0.016	-0.081**	0.008	-0.033	0.005	-0.032	0.006	-0.025
College GPA	0.104***	0.055***	0.087***	0.031**	0.088***	0.022	0.068**	0.014
High school GPA	0.01	0.057***	0.01	0.042**	0.01	0.034*	0.012	0.037
Student experiences			0.182***	0.013	0.169***	0.014	0.157***	0.011
<i>Participation in study abroad program</i>								
<i>Studying with other students</i>			0.064**	-0.048**	0.056**	-0.039*	0.046*	-0.028
<i>Presentation of ideas and perspectives</i>			-0.094***	0.038**	-0.078**	0.035*	-0.06*	0.032
Discussion with other students in class			-0.067**	0.095***	-0.081**	0.09***	-0.076**	0.08***
<i>Participation in the setting of class themes</i>			0.065**	0.087***	0.06**	0.075***	0.024	0.062***
<i>Research of literature and materials</i>			-0.009	0.092***	-0.025	0.076***	-0.052*	0.059***
<i>Studying and working on assignments outside of class</i>					0.029	0.052**	-0.007	0.044*
Attending classes and experiments					0.039	0.022	0.055**	0.027
Reading (novels and others)					-0.015	-0.023*	-0.035	-0.028
Part-time work outside university					0.088***	0.073***	0.068***	0.066***
<i>Providing advice on in-class and academic work</i>							0.032	0.075***
Writing recommendation of letter							0.054*	-0.039*
Providing emotional support and encouragement							0.117***	0.042*
Fixed	18.755	16.053	19.354	13.802	18.638	13.231	16.853	12.795

R square	0.042	0.049	0.11	0.097	0.116	0.105	0.148	0.115
N	2282	4177	2236	4159	2179	4131	2112	4127
Outcome: human personal skills	Japan	Korea	J	K	J	K	J	K
Attribute	b	b	b	b	b	b	b	b
<i>Gender dummy</i>	-0.064**	0.121***	-0.027	0.079***	-0.015	0.073***	-0.013	0.055***
Entrance exam dummy	0.005	-0.027	-0.012	-0.025	-0.011	-0.03*	-0.001	-0.031*
Humanities dummy	-0.058**	-0.106***	-0.066**	-0.089***	-0.06**	-0.094***	-0.045	-0.064**
Social sciences dummy	-0.065**	-0.063*	-0.043	-0.029	-0.038	-0.025	0.002	0.019
STEM dummy	-0.135***	-0.071**	-0.114***	-0.019	-0.11***	-0.025	-0.088**	0.001
<i>College GPA</i>	0.134***	0.107***	0.114***	0.074***	0.111***	0.068***	0.111***	0.051**
High school GPA	0.051**	-0.005	0.035	-0.017	0.039	-0.02	0.042*	-0.012
<i>Participation in study abroad program</i>			0.053**	-0.007	0.054**	-0.005	0.053**	-0.01
Student experiences								
Studying with other students			0.198***	-0.209***	0.182***	-0.199***	0.151***	-0.175***
Presentation of ideas and perspectives			-0.096***	0.108***	-0.085***	0.098***	-0.051*	0.093***
Discussion with other students in class			-0.076**	0.064***	-0.073**	0.066***	-0.052*	0.047**
Participation in the setting of class themes			0.015	0.061***	0.021	0.06***	-0.005	0.031
Research of literature and materials			0.072***	0.05**	0.069**	0.038*	0.04	0.003
<i>Studying and working on assignments outside of class</i>					0.088***	0.08***	0.042	0.062***
Part-time work outside university					0.073***	-0.029	0.079***	-0.029*
<i>Providing advice on in-class and academic work</i>							0.045	0.079***
Faculty engagement							0.089**	0.036

(continued)

Table 14.4 (continued)

		Model 1		Model 2		Model 3		Model 4	
Outcome: Global competencies		Japan	Korea	J	K	J	K	J	K
		b	b	b	b	b	b	b	b
	Encouragement to pursue graduate or professional study							0.105***	0.051*
	<i>Providing feedback on your academic work</i>							0.023	0.063***
	<i>Providing opportunities to apply class experience to real life</i>							0.04	-0.07***
Fixed		17.101	22.336	16.487	21.274	15.733	20.848	13.942	20.256
R square		0.045	0.026	0.143	0.122	0.155	0.129	0.199	0.157
N		2290	4179	2245	4162	2186	4134	2120	4130
Outcome: Cognitive ability		Japan	Korea	J	K	J	K	J	K
Attribute		b	b	b	b	b	b	b	b
	<i>Gender dummy</i>	-0.01	0.122***	0.023	0.076***	0.019	0.075***	0.021	0.066***
	Entrance exam dummy	0.017	0.034*	-0.004	0.029*	-0.027	0.018	-0.015	0.015
	Humanities dummy	-0.054*	-0.048	-0.06**	-0.035	-0.043	-0.041	-0.039	-0.019
	Social sciences dummy	-0.02	-0.097***	-0.001	-0.06*	0.012	-0.055*	0.042	-0.023
	STEM dummy	-0.062*	-0.117***	-0.048*	-0.048	-0.045	-0.066**	-0.037	-0.042
	College GPA	0.156***	0.141***	0.134	0.104***	0.112***	0.095***	0.107***	0.086***
	High school GPA	0.01	0.058***	-0.002	0.041**	-0.005	0.03	-0.007	0.034*
Student experiences	Participation in study abroad program			0.036	-0.04**	0.035	-0.035*	0.027	-0.033*

For both Japanese and Korean students, the students' overall satisfaction with their university experience had a positive influence on *interpersonal skills*. Although studying with other students had a positive influence on *interpersonal skills* among Japanese students, the impact was negative for Korean students. For the Japanese students, although attending class and experiments did not have a positive influence on *interpersonal skills*, off-campus part-time or full-time work did. For Korean students, study time outside of class had a positive influence on *interpersonal skills*. In addition, for both Japanese and Korean students, reading books did not have a positive influence on *interpersonal skills*. Faculty engagement had a greater influence, either negative or positive, on the *interpersonal skills* of Korean students than that of Japanese students. With respect to input (I) factors, for Japanese students, both high school and university GPA had a positive influence on *interpersonal skills*. The influence was greater for students in health- and medicine-related majors than for students in any other majors. For Korean students, entrance into university by means other than the general entrance exam resulted in greater improvement of *interpersonal skills*.

For both Japanese and Korean students, the students' overall satisfaction with their university experience had a positive influence on *cognitive ability*, and students with higher GPAs had a greater tendency to report improvement of *cognitive ability*. With respect to students' experience, the predictive factors for improvement of *cognitive ability* were more similar between Japanese and Korean students than for other learning outcomes. For both Japanese and Korean students, study time outside of class, attendance of classes, and reading had a positive influence on *cognitive ability*, while on-campus part-time or full-time work had a negative influence. Although more variables related to students' experience had a significant influence (either positive or negative) on this outcome for both Japanese and Korean students, the magnitude of the influence was greater for Korean students. While the influence of "presenting thoughts and research" on *cognitive ability* was negative for Japanese students, the impact was positive for Korean students. For Korean students, whereas "discussion among students," "deciding class themes," and "research of literature and other materials" had a positive influence on *cognitive ability*, "study abroad programs" and "studying with other students" did not, indicating that experience with active learning has a strong impact for Korean students. Whereas individual faculty engagement in the form of "encouragement to pursue graduate or professional studies" and "emotional support" had a positive influence on *cognitive ability* for Japanese students, "instructors' advice on students' class and academic work" had a negative influence on *cognitive ability* for Korean students.

14.5.5 Construction of Student Types

Principle component analysis of 16 items related to self-reported abilities and skills or behavioral characteristics using varimax rotation yielded three components (factor loading, 0.400 or greater; cumulative contribution ratio, 54.56%), which I

named *proactive behavior characteristic*, *empathy characteristic* and *cognitive characteristic*.⁴ Assessment of component reliability using Cronbach's alpha revealed the components to be reasonably reliable (α of 0.826, 0.800, and 0.696, respectively).

Next, cluster analysis based on the scores for each component using Ward's method resulted in the identification of five student types shown in Fig. 14.2. With respect to the characteristics of each student type, Type 1 students have high *proactive behavior characteristic* and *cognitive characteristic* scores and low *empathy characteristic* scores. The image that emerges is of a student that is highly motivated and approaches things with confidence and a can-do attitude, has confidence in his or her own *cognitive ability*, is not proficient at understanding others or oneself, and does not have much confidence in expressing him or herself in writing. Type 2 students have low scores for all components. The image that emerges is of a student who has less confidence in his or her ability, skills, and actions than his or her peers. This student type constitutes the largest share of students in the sample (2665 students, 34.6%). Type 3 students have low *proactive behavior characteristic* and *cognitive characteristic* scores but much higher *empathy characteristic* scores compared to other student types, suggesting that these students are very confident in their ability to develop and maintain interpersonal relationships. Type 4 students have high *empathy characteristic* and *cognitive characteristic* scores. As such, the image that emerges is of a student who does not act proactively but who understand themselves well and gets along well with others and has confidence in their *cognitive ability*. Type 5 students have low *cognitive characteristic* scores. As such, the image that emerges is of a student who acts proactively and is good at developing interpersonal relationships but who has relatively lower confidence in his or her *cognitive ability*. However, it should be kept in mind that these are results of analysis based on component scores and, therefore, that a low score for any component means that the student has lower confidence relative to the entire sample population but does not mean that the student has no confidence at all.

⁴The variables contributing to proactive behavior characteristic include [stability of mood], [physical health], [can-do spirit], [leadership], [motivation], and [presentation skills]. The variables contributing to empathy characteristic include [spirituality], [self-understanding], [writing ability], [understanding of others], and [confidence in social skills]. The variables contributing to cognitive characteristic include [mathematical ability], [IT skills], [academic achievement], and [confidence in intellectual ability].

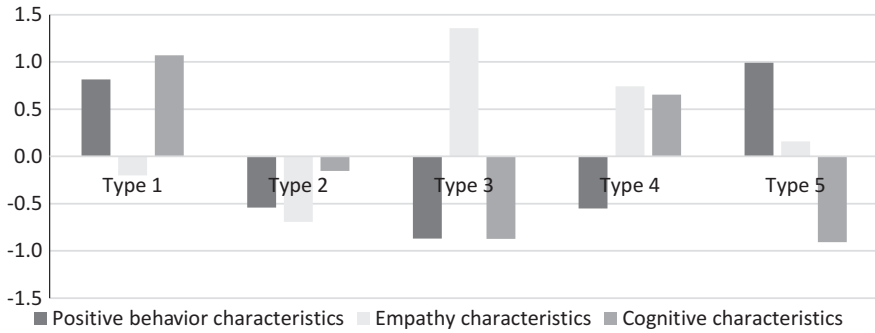


Fig. 14.2 Student typology based on proactive behavior, empathy, and cognitive characteristics
Note: The numbers of students assigned to each type are as follows—Type 1 ($n = 1486$), Type 2 ($n = 2665$), Type 3 ($n = 692$), Type 4 ($n = 1296$), and Type 5 ($n = 1569$)

Table 14.5 compares the frequency of negative experiences of Japanese and Korean students broken down by student typology based on the students' self-assessment of their abilities, skills, and behavioral characteristics. Type 2 students, who have relatively less confidence in terms of all three characteristics, constitute the largest group of Japanese students attending private universities. The next largest group is Type 5 students, who have high *proactive behavior characteristic* scores, average *empathy characteristic* scores, and low *cognitive characteristic* scores. In contrast, student types are more evenly distributed among Korean students attending private universities. That said, the largest group is Type 2 students, who have low confidence in terms of all three characteristics, followed by Type 1 students, who have low confidence in *empathy characteristic* scores, and Type 4 students, who have low *proactive behavior characteristic* scores. Type 3 students constitute the smallest group of students in both Japan and South Korea.

Next, comparing means scores for negative experiences between Japanese and Korean students broken down by student type (Table 14.5), the only item for which the mean score is higher for Japanese students than for Korean students across all student types is "I felt the class was boring." Scores did not differ substantially among student types. For all other items, mean scores were higher for Korean students than Japanese students across all student types. The greatest differences between Japanese and Korean students were observed for the two items "I could not complete the homework by the deadline" and "I asked for counseling." While study time outside of class and self-reports of improvement of learning outcomes were also higher for Korean students than for Japanese students, these results are perhaps an indication of the pressure felt by Korean students in a highly competitive environment in which the students perceive of counseling as an everyday service in Korean private universities.

Table 14.5 One-way ANOVA of negative experiences of Japanese and Korean students

Type of students	Country	I was unable to complete my homework on time	I felt that bored in class	I was late for class	I had to miss class due to work	I felt sad	I was overwhelmed by all that I had to do	I received counseling
1	Japan (N = 352)	1.28	1.90	1.66	1.32	1.49	1.79	1.12
	Korea (N = 1134)	2.46	1.89	2.25	2.29	1.75	2.06	2.37
2	Japan (N = 1239)	1.37	1.99	1.72	1.23	1.66	1.85	1.09
	Korea (N = 1423)	2.42	1.73	2.21	2.27	2.00	2.20	2.52
3	Japan (N = 230)	1.38	2.04	1.84	1.42	1.96	2.07	1.17
	Korea (N = 462)	2.41	1.67	2.17	2.19	2.05	2.37	2.44
4	Japan (N = 279)	1.34	1.86	1.72	1.33	1.69	1.89	1.18
	Korea (N = 1017)	2.45	1.77	2.21	2.24	2.01	2.27	2.41
5	Japan (N = 701)	1.38	2.02	1.81	1.29	1.55	1.92	1.06
	Korea (N = 866)	2.44	1.79	2.23	2.28	1.80	2.16	2.45
	Total (N = 7703)	2.05	1.85	2.05	1.91	1.81	2.08	1.95
	F value	39.232 ($p < 0.00001$)	8.016 ($p < 0.00001$)	10.241 ($p < 0.00001$)	42.682 ($p < 0.00001$)	52.55 ($p < 0.00001$)	28.362 ($p < 0.00001$)	42.923 ($p < 0.00001$)

1, not at all; 2, sometimes; 3, often

14.6 Conclusions

In this chapter, I compared the relationships between students' experience and learning outcomes, differences in study time, and factors predicting learning outcomes for Japanese and Korean student in different areas-of-study. In addition, I developed student typologies based on students' self-evaluations of their behavioral characteristics and analyzed the relationships between student type and negative experiences in school. The analyses yielded the following four conclusions:

First, experience of active learning had a positive, albeit small, effect on learning outcomes for both Japanese and Korean students. While several differences in skills acquisition were observed among Japanese students in different areas-of-study, the magnitude (and frequency) of such differences was not as great for Korean students as it was for Japanese students.

Second, study time outside of class was found to differ between Japanese and Korean students. With the exception of students in health- and medicine-related majors, Korean students attending private universities spent more time studying outside of class than their Japanese counterparts. However, it is not possible to break down this study time further, as the data reflect students' self-reports of their study time from a single question on the questionnaire. Kaneko (2013) provides rich insight into the breakdown of study time. Although Kaneko's "autonomous study" is not directly constrained in terms of time or space, Kaneko defines "autonomous study" as study that is carried out at a time and place and in a manner determined by the student him or herself within the educational framework provided by the university. This study includes time spent completing assignments, preparing for or reviewing classes, as well as carrying out research and writing graduation theses.

That said, currently, many universities are introducing elements of active learning, which I discuss in greater detail below, and opportunities for students to engage in peer learning, whereby students study together in groups, are increasing. It is not clear whether the students considered such peer learning as part of their study time and included it in their estimates of study time when responding to the questionnaire. It is possible that Japanese student consider study time to only include times spent studying alone. Accordingly, the students' responses to the question on study time may be influenced by this mindset. Thus, this study is limited in its ability to examine the structure of study time.

Third, for both Japanese and Korean students, time spent on various activities including study time outside of class, faculty engagement, and experience with active learning are all predictive factors for the three learning outcomes investigated in this study, namely, *global competency*, *interpersonal skills*, and *cognitive ability*. Particularly in the case of Korean students, experience with active learning has a similar substantial positive impact on the acquisition of skills for students in all disciplines. The structure of Japanese STEM curricula is such that it is difficult to systematically incorporate classes dealing with current issues or elements of active learning. In contrast, many Korean universities have recently implemented changes to push back the timing with which students choose their areas-of-study (i.e.,

encourage “late decision”).⁵ I would note that the fact that many Korean students take the same general and liberal arts in their first 2 years of university may contribute to the smaller difference between disciplines observed in this study.

Fourth, I believe that the relationship between student typology and negative experiences elucidated by this study provides insight into to how Japanese university environments can or should be changed in the future. A relatively high number of Japanese students enrolled in private universities have low confidence in the three behavioral characteristics identified in this study. The challenge for universities is to create environments that raise the confidence of such students. Meanwhile, while fewer Korean students are of the type that has relatively low confidence overall, students of all types tend to have more negative experiences than their Japanese counterpart. Although I was unable to determine specific characteristics of the environment of Korean universities in this study, one direction of reform would be to achieve a better balance between students’ academic and non-academic experiences. As such, I plan to further analyze this data in relation to students’ academic and non-academic experiences.

References

- Choi, J., & Rhee, B. (2009). Examining factors related to college students’ learning outcomes: Focusing on effects of college. *The Journal of Educational Administration*, 27(1), 199–222.
- Choi, J., Kim, M., Yi, P., & Lee, E. (2009). *Research on the strategic participation in OECD AHELO project for enhancing the higher education competitiveness*. Korean Educational Development Institute.
- Furuta, K. (2010). Learning environment of university and learning outcomes: Self-reported student data to examine acquisition of knowledge and skills. *Journal of Quality Education*, 3, 59–75.
- Kim, A., & Rhee, B. (2003). An analytic study of identifying personal and institutional influences on the perceived development of core competencies of college students. *The Journal of Korean Education*, 30(1), 367–392.
- Matsushita, K. (2014). Competences as learning outcomes and their assessment: Potential and challenges of rubric-based assessment. *Nagoya Journal of Higher Education*, 14, 235–256.
- Mizokami, S. (2009). Student learning and development from a college life perspective: Well-balanced curricular and extra-curricular activities show high development. *Kyoto University Studies in Higher Education*, 15, 107–118.
- Morozumi, A. (2009). How student’s learning behavior differs among universities?: Focusing on the effect of mode of teaching. *Bulletin of the Graduate School of Education, the University of Tokyo*, 49, 191–206.
- OECD. (2012). *Education at a glance 2012 highlights*. Retrieved November, 21, 2014, from <http://www.oecd.org/edu/highlights.Pdf>
- Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research*. San Francisco: Jossey-Bass.
- Pike, G. R., & Kuh, G. D. (2005). A typology of student engagement for american college and universities. *Research in Higher Education*, 46(2), 185–209.

⁵Based on reports by Korean researchers.

- Rhee, B. S. (2013). *Gains in learning outcomes of Korean and Japanese college students: Factors affecting the development of generic skills in undergraduate students*. Paper presented at AIR2013, Long Beach.
- Sugitani, Y., Yamada R., & Yoshida A. (2013). *Learning and Engagement of College Students based on the Field of Study*. Presentation material of 64th Japan Society of Educational Sociology.
- Tanimura, H. (2010). Learning hours and learning outcomes of college students. *The Journal of Management and Policy in Higher Education*, 1, 71–84.
- Yamada, R. (2012). *Quality assurance for undergraduate education: Based on the results of student survey and first-year experiences*. Tokyo: Toshindo.
- Yamada, R. (2014a). Gains in learning outcomes of Korean and Japanese college students: Based on cases of Japanese students. *The International Education Journal: Comparative Perspectives*, 13(1), 100–118.
- Yamada, R. (2014b). Students and active learning: How the learning commons support student's learning. *Journal of the Liberal and General Education Society of Japan*, 36(1), 32–40.