3.5 Measuring Scientific Reasoning Competencies

Multiple Aspects of Validity

Krüger, D., Hartmann, S., Nordmeier, V., and Upmeier zu Belzen, A.

Abstract

In this chapter, we investigate multiple aspects of validity of test score interpretations from a scientific reasoning competence test, as well as aspects of reliability. Scientific reasoning competencies are defined as the disposition to solve scientific problems in certain situations by conducting scientific investigations or using scientific models. For the purpose of measurement, the first phase of our project focused on the construction of a paper-pencil assessment instrument – the *KoWADiS* competence test – for the longitudinal assessment of pre-service science teachers' scientific reasoning competencies over the course of academic studies. In the second phase of our project, we investigated the reliability of the test scores and the validity of their interpretations. We used a multimethod approach, addressing several sources of validity evidence. Overall, the results are coherent and support the validity assumptions to a satisfactory degree. The long-term goal is the use of this test to provide empirically sound suggestions for pre-service science teacher education at university level.

The electronic edition of this chapter has been revised: The university name for the authors S. Hartmann and A. Upmeier zu Belzen has been corrected. A Correction is available at https://doi.org/10.1007/978-3-658-27886-1_21

O. Zlatkin-Troitschanskaia et al. (Hrsg.), *Student Learning in German Higher Education* https://doi.org/10.1007/978-3-658-27886-1_13

Keywords

Scientific reasoning competencies, teacher education, validity

1 Introduction

Scientific reasoning competencies are a set of acquired cognitive dispositions that individuals such as scientists, teachers, or students use to solve scientific problems systematically. The assessment of cognitive dispositions is possible by measuring manifest behavior, i.e. performance (Koeppen et al. 2008). Scientific reasoning competencies are performed by applying skills like planning, conducting, and evaluating scientific investigations, or using scientific models (Fischer et al. 2014). These inquiry processes allow to gain new insights into scientific phenomena, utilizing research questions, theories and hypotheses as typical aspects of the hypothetico-deductive approach of the empirical sciences (Popper 2004). Competence tests as construct-related measurements are needed to assess what students are able to do as a result of their learning (Blömeke et al. 2015; Osborne 2013). Competence tests indicate students' performance based on reliably and validly interpretable criterion-related measures. As sources of evidence for validity we investigated test content, response processes, internal structure, and relations to other variables (AERA et al. 2014). These criteria serve as sources of evidence and were applied systematically to test our instrument, the *Ko-WADiS* competence test for scientific reasoning (Hartmann et al. 2015a; Mathesius et al. 2014; Stiller et al. 2016; Straube 2016).

In this chapter, we present evidence for the validity and reliability of our test score interpretations, discuss implications for the theoretical model and for the test instrument and its practical use, and provide an outlook for further use of the model and the test.

2 Theoretical Framework: Scientific Reasoning Competencies

Scientific reasoning (Giere et al. 2006; Klahr 2000) as a problem-solving process (Mayer 2007) is considered a key competence in basic science education for the natural sciences biology, chemistry, and physics (Rönnebeck et al. 2010, p. 178). As such it belongs to the indispensable core competencies for the 21st century (Trilling and Fadel 2009). Scientific reasoning competencies are cognitive dispositions

to gain empirical insights into scientific phenomena by successfully applying steps of an idealized problem-solving process to given scientific problems (Gut-Glanzmann and Mayer 2018; Mayer 2007). Within the three empirical natural sciences biology, chemistry and physics, central methods are scientific observation of phenomena, controlled experimentation by the variation and control of variables (Gut-Glanzmann and Mayer 2018; Mayer 2007; Wellnitz and Mayer 2013), as well as scientific modeling (Krüger et al. 2018; Upmeier zu Belzen and Krüger 2010).

In criterion-driven observations, competencies in predicting, describing, and systematically examining correlative relationships between structures and their functions under temporal changes are required (Wellnitz and Mayer 2013). Competencies in systematic experimentation reflect the ability to capture causal relationships with systematically varied and controlled conditions (Gut-Glanzmann and Mayer 2018; Mayer 2007). This requires the ability to handle independent and dependent variables as well as control variables (Roberts and Gott 2003). According to Mayer (2007), the scientific thinking processes underlying observations and experimentation as scientific investigations can be described as a domain-specific form of problem solving in four sub-competencies, which were operationalized in the Ko-WADiS competence test (Table 1): Formulating research questions, generating hypotheses, planning investigations, and analyzing and interpreting data (Gut-Glanzmann and Mayer 2018). These sub-competencies generally refer to experimentation, but can also be applied to observations, comparisons (Wellnitz and Mayer 2013), and the use of models (Upmeier zu Belzen and Krüger 2010). However, there are specific reasoning competencies needed in scientific modeling. Scientists, students, and teachers need to be able to reflect the role of models in the process of scientific modeling (Krüger et al. 2018). Thus, it has to be assessed whether and to which extent models are seen as tools to reconstruct central features of reality or to develop a methodological basis for the formulation of research questions (Gilbert and Justi 2016). According to Upmeier zu Belzen and Krüger (2010), using scientific models to reason about scientific phenomena is operationalized in sub-competences (Table 1) such as purpose of models, testing models, and changing models.

These seven sub-competencies of conducting scientific investigations and using scientific models (Table 1) are interrelated steps of a general scientific thinking process in the sense of the hypothetico-deductive approach (Popper 2004). Against the background of an ideal-typical view, a researchable scientific question is formulated on the basis of a real-life scientific problem (White 2017). Subsequently, a model is developed whose purpose is to generate inter-subjectively traceable and falsifiable hypotheses (Lawson et al. 2000). Testing these hypotheses means testing the model. For this, a suitable experimental arrangement must be planned

(Lawson et al. 2000). The results must be evaluated and interpreted and can either support or oppose the assumptions of the model. The latter option results in changing the model, which means that the process restarts. Being competent in the field of scientific reasoning is defined as a cognitive disposition that enables students to apply each of these seven steps to real-life scientific problems (Giere et al. 2006).

Table 1 Scientific reasoning competencies in the areas *conducting scientific investigations* and using *scientific models*

Competency	Scientific reasoning	
Dimension	conducting scientific investigations	using scientific models
Sub-competencies	formulating questions (18)	purpose of models (18)
	generating hypotheses (16)	testing models (18)
	planning investigations (22)	changing models (14)
	analysing data and drawing conclusions (17)	

Note. Number of developed items in brackets

The described sub-competencies become accessible to measurement by the cognitive-psychological construct of scientific reasoning (Fischer et al. 2014). In international research, they are also conceptualized in styles of scientific reasoning (Osborne 2018) or scientific paths of knowledge acquisition (Priemer et al. 2019). In science curricula (e.g., KMK 2014; NGSS 2013), the ability to carry out and reflect about scientific inquiry processes is mandatory.

3 The Ko-WADiS Competence Test

In the project, *Ko-WADiS*[1] (2011–2015), the paper-pencil *Ko-WADiS* competence test was developed (Hartmann et al. 2015a). The focus was on the clarification of the theoretical foundations, the development of the test instrument, the standardization of the items for the three subjects biology, chemistry, and physics (Mathesius et al. 2014), the investigation of the basic psychometric properties of the test, and the start of a longitudinal assessment in the target population (pre-service science teachers) as well as in various control groups (Hartmann et al. 2015a).

1 **Ko**mpetenzmodellierung und -erfassung zum **W**issenschaftsverständnis über naturwissenschaftliche **A**rbeits- und **D**enkweisen be**i** **S**tudierenden

The *Ko-WADiS* competence test is the result of a multi-step process (Hartmann et al. 2015b), in which students' responses in the open-ended format were checked for content validity in an expert discourse. The piloting of the multiple-choice items finally led to a set of 123 items (distribution see Table 1). To make the data collection and evaluation more economical, and to reduce the amount of missing values, 63 items with the best psychometric properties were selected from this item pool (three items per subject and sub-competence; Stiller et al. 2016). These items are used for longitudinal assessment since 2013. The paper-pencil-based single best answer items address scientific reasoning competencies as they are typical for pre-service teachers of biology, chemistry and physics (Hartmann et al. 2015a), but were also used by science students, students of psychology, and in-service biology teachers. Each test booklet contains 21 of the 63 items (balanced-incomplete block design; Gonzalez and Rutkowski 2010), and is assessed in 45 minutes. In each test booklet, the seven sub-competences and the three scientific disciplines are equally distributed. Because of the design, the test is evaluated using probabilistic methods.

4 Investigation of Validity

The validity of the *Ko-WADiS* test score interpretations was evaluated in a follow-up project, *ValiDiS*[2] (2016–2019). Alongside a continuation of the longitudinal data collection to investigate competence development, we addressed different sources of validity evidence as described in the *Standards for Educational and Psychological Testing* (AERA et al. 2014; see also Kane 2013): evidence based on test content, evidence based on internal structure, evidence based on response processes, and evidence based on relations to other variables such as conceptually related constructs and criteria. Investigations on these sources of validity provide empirical evidence to support the assumption that the test results can be interpreted in terms of the underlying theoretical construct.

Finally, consequences of testing can serve as a potential source of validity evidence, but it was not used in our project as no consequences suitable for validity investigations are based on the test results yet.

2 **Ko**mpetenzmodellierung und -erfassung: **Vali**dierung eines Modells zum wissenschaftlichen **D**enken **i**m naturwissenschaftlichen **S**tudium

4.1 Evidence Based on Test Content

An investigation of content validity was addressed from the beginning of the *Ko-WADiS* project, thus starting prior to the development of the test instrument. To ensure a standardized item construction process which followed guidelines based on theoretical groundings (Mayer 2007; Upmeier zu Belzen and Krüger 2010), an item construction manual was developed. Answering options for the single-best answer items of the developed instrument were based on written answers of students to open-ended items (Hartmann et al. 2015b; Mathesius et al. 2014). Investigations of item features that systematically affect item difficulty revealed predictive potential for one formal item feature (length of response options), two features based on cognitive demands (processing data from tables, processing abstract concepts), and one feature based on solid knowledge (specialist terms). This was in accordance with the cognitive demands operationalized in the theoretical structure of the test. Thus, it is concluded that the findings support the validity of the interpretation of the test scores as measures of scientific reasoning competencies (Stiller et al. 2016).

Content validity was also examined by an expert rating. A sample of 21 academic teachers and researchers with a high level of theoretical knowledge and research experience in the field of scientific reasoning (Gruber 2010) were requested to classify the relationship between six selected items (two per scientific discipline) and the sub-competencies of the theoretical construct (Table 1). The experts correctly designated each item to the corresponding sub-competence, and evaluated how well the item represented the sub-competence on a five-point Likert scale with a value 1 for very poor and a value of 5 for very good theoretical resemblance of the item. The median rating was 4 out of 5 for five items, and 3 out of 5 for one item, with an interquartile range between 0.00 and 2.25. These results indicate an appropriate operationalization of the items, indicating that they represent the construct to a satisfactory degree (Hartmann et al. 2019a).

4.2 Evidence Based on Internal Structure

With respect to the internal structure, the empirical results from cross-sectional data support a one-dimensional structure of scientific reasoning, although the fit of a two-dimensional model that differentiates between aspects of scientific investigation and aspects of scientific modeling is not significantly worse than that of the one-dimensional model (Hartmann et al. 2015a). This is in accordance with the theoretical assumptions of the underlying construct, which assumes scientific

reasoning being generalizable across the subjects biology, chemistry and physics. Thus, the established unidimensionality indicates a broad theoretical construct. These results correspond with those of other empirical studies on scientific reasoning competencies (e.g., Mannel 2011; Neumann 2011; Wellnitz 2012).

4.3 Evidence Based on Response Processes

Response options (distractors and attractors) were generated out of students' written answers to open-ended questions, using the generated item stems as stimuli. This contributes to student-centered response options in the single best answer test and secures valid test score interpretations.

The investigation of response processes as a source of validity evidence was done by eyetracking collecting gaze data and verbal data while working on items of the Ko-WADiS competence test (12 items, $N = 16$; Mathesius et al. 2018). In addition to think-aloud protocols, the cued retrospective reporting method (attention maps, sequence charts; van Gog et al. 2005) was applied for the investigation of cognitive processes during eye tracking. Although pre-service biology teachers who chose the attractor do not differ in their eye movement patterns (fixations, dwell time) from those who chose the distractor, the verbal data describes the individual solution processes in a comprehensible way. The findings based on response processes are interpreted as evidence for the validity of the test scores interpretation as measures of scientific reasoning competencies (Mathesius et al. 2018).

4.4 Evidence Based on Relations to other Variables

Validity evidence based on relations to other variables was investigated utilizing several empirical methods (Table 2). An investigation of instructional sensitivity was accomplished by observing the development of test scores due to short-term learning progress in regular university seminars and lectures (1) and in intervention studies (2). Besides that, we investigated groups with either hypothesized mean differences (3) or mean equivalence (4). Finally, correlations with general abilities like intelligence and complex problem solving (5) and with conceptually related constructs, such as pre-service teachers' pedagogical content knowledge (6), were calculated.

1. Instructional sensitivity is supported by an investigation during regular academic training in courses of biology education: We used the long version of our instrument to test 59 pre-service biology teachers before and after a semester. Comparing the results, we found a moderate increase in the students' ability estimates ($t = 2.59$, $_{pone\text{-}sided} = .006$, $d = 0.34$). Instructional sensitivity was also investigated in an interventional study. A sample of 87 pre-service science teachers participated on a two-day intensive course to train scientific reasoning competencies. The instructional sensitivity was tested for a selection of nine items from our original instrument which were answered by the students before and after the intervention. The pre-post comparison of the sum score reveals a significant increase ($t = 2.30$, $p_{one\text{-}sided} = .012$, $d = 0.25$). In a control group ($N = 55$), no significant effect was found.

   A sample of 125 pre-service biology teachers participated on a seminar to promote scientific reasoning competencies explicitly. The instructional sensitivity was tested with an item subset of 21 biology items before and after the intervention. The pre-post comparison of the sum score reveals a significant increase ($t = 4.72$, $p_{two\text{-}sided} < .001$, $d = 0.35$). In a control group ($N = 49$), no significant effect was found. These findings further support the interpretation of the test scores as measures of scientific reasoning competencies.
2. Known-group comparisons (Cronbach and Meehl 1955) are an economical tool to investigate aspects of criterion-based validity. We used it to test whether our test scores are sensitive to differences between undergraduate and postgraduate students, and to differences between pre-service science teachers who study one scientific discipline alongside a non-scientific discipline and pre-service science teachers who study two scientific disciplines. The comparisons were carried out as a latent regression analysis. The results support the hypotheses of group differences with significant regression effects of group affiliation on the latent ability measures ($B_{academic\ phase} = 0.283, p < .001$; $B_{scientific\ disciplines} = 0.116$, $p < .01$).

   Additionally, known-group comparisons were carried out with 626 pre-service biology teachers (Mathesius et al. 2016). It was predicted that pre-service biology teachers who also study chemistry or physics perform better than pre-service biology teachers without a second science subject, and pre-service biology teachers in more advanced stages of academic education (study stages: 4th-7th semester and 8th-10th semester) perform better in the test than students in early stages (1st-3th semester). To test these hypotheses, multiple latent regression analysis was applied. The results show significant regression effects of group affiliation ($B_{scientific\ disciplines} = 0.774, p < .001$; B_{terms} 4–7 $= 0.499, p < .001$; B_{terms} 8–10 $= 1.279, p < .001$) on the latent ability measures. Both findings indicate

that the test scores are sensitive to relevant criteria (Hartmann et al. 2015b; Mathesius et al. 2016).

3. In addition to the study of predicted mean differences, predicted mean equality was tested as well (Hartmann et al. 2019b). On the basis of initial grades, course and module descriptions, it was hypothesized that the levels of pre-service science teachers' and psychology students' scientific reasoning competencies do not show a meaningful difference. Therefore, the mean test scores in the two groups should be equivalent. To test the hypothesis, we compared the means of matched sub-samples with balanced covariate distributions, utilizing the two-on-one-sided-tests method (TOST; Schuirmann 1987) to test the equivalence of the students' abilities. The hypothesis of group equivalence is supported by the absence of a significant difference ($t = 0.03$; $p_{two\text{-}sided} = .98$) in combination with a very small effect size of d = 0.00 that is nominally below a pre-defined smallest effect size of interest (d = 0.17). However, the TOST procedure indicates that the confidence interval around the mean difference exceeds the equivalence bounds due to the relatively small sample size, rendering the results inconclusive ($t_{TOST} = 1.35$; $p_{one\text{-}sided} = .09$; Hartmann et al. 2019b).
4. As a further indicator of validity, it was tested to what degree variance of the *Ko-WADiS* test scores can be attributed to more general skills such as intelligence or complex problem-solving abilities (Mathesius et al. 2019). The *Ko-WADiS* competence test scores and the scores of the reasoning scale of the Intelligence-Structure Test 2000 R (Liepmann et al. 2007) and complex problem solving (Genetics Lab test; Greiff and Fischer 2013; Sonnleitner et al. 2013) show positive significant correlations (I-S-T 2000 R: $r = .44$; $p_{two\text{-}side}\ d < .001$; Genetics Lab test: $r = .33\text{-}.40$; $p_{two\text{-}sided} < .001$). Furthermore, the regression analysis clarifies 24 % of the variance by the considered variables. The findings support the assumption that they are distinct constructs with connecting facets. Therefore, part of the remaining variance might be interpreted as evidence for the test score interpretations as measures of scientific reasoning competencies (Mathesius et al. 2019).
5. Finally, the convergent validity was investigated in a correlational study ($N = 65$ pre-service science teachers). We correlated sum scores from a short version of our instrument (15 items) with 12 scientific-reasoning items from an early testing version of the PCK-IBI (Großschedl et al. 2018). The scores from the two instruments correlate significantly ($r = .49$; $p_{one\text{-}sided} < .001$). Potential moderating effects of general cognitive ability were controlled by including the non-verbal subscale of the IST Screening (Liepmann et al. 2007) in the analysis, which left the correlation coefficient practically unchanged ($r_{partial} = .48$; $p_{one\text{-}sided} < .001$). The findings indicate that the correlation of the two tests is not explained by

intelligence, which provides further supporting evidence to the validity of our test score interpretations as measures of scientific reasoning competencies.

4.5 Summary

Overall, the outcomes of the validation studies are coherent and provide supporting evidence for the interpretation of the test scores as measures of scientific reasoning competencies. The majority of the investigated effects were small to medium, implying that the true effects are moderate. However, statistical power is limited due to mediocre reliabilities (Section 5), which certainly affect the power of the statistical procedures (Kanyongo et al. 2007).

Table 2 Overview of studies for sources of validity evidence

Source of validity evidence	Investigation of …	Instrument and sample	Results	Reference
Test content	… process of item development	183 open ended items $N = 259$ 166 single best answer items $N = 578$	theory-based selection of 123 items based on item-parameter analysis	Hartmann et al. 2015b
	… item features affecting item difficulty	63 single best answer items $N = 907$; 9 experts	32 % of variance is explained by the item features investigated, and is in accordance with model assumptions and expert ratings from a standard setting	Stiller et al. 2016
	… expert ratings of selected test items	6 Likert-style items, 21 experts	selected items represent the theoretical construct to a satisfying degree	Hartmann, Krüger et al. 2019
Internal structure	… the dimensionality of the theoretical structure	141 items, $N = 3\,010$	unidimensional structure in accordance with theoretical assumptions	Hartmann et al. 2015a; Hartmann et al. 2015b

Source of validity evidence	Investigation of …	Instrument and sample	Results	Reference
Response processes	… gaze data and think aloud protocols	12 single best answer items $N = 16$	no correlation between selection of answer and eye movements, scientific reasoning is necessary to find the attractor	Mathesius et al. 2018
Relations to other variables	… the instructional sensitivity to progress in selected seminars and lectures	123 single best answer items, pre-post design, $N = 59$	moderate increase of scientific reasoning competencies ($d = 0.30$)	
		21 single best answer biology items, pre-post design, $N = 49$	increase of scientific reasoning competencies ($d = 0.54$)	Mathesius et al. in preparation
	… the instructional sensitivity in intervention studies	9 single best answer biology items, $N = 87$	increase of scientific reasoning competencies ($d = 0.25$)	Hartmann, Krüger et al. 2019
		21 single best answer biology items, $N = 125$	increase of scientific reasoning competencies ($d = 0.77$)	Mathesius et al. in preparation
	… groups with hypothesized mean differences	123 single best answer items, $N = 2247$	significant regression effects support hypothesized group differences	Hartmann et al. 2015b
		123 single best answer items, $N = 626$ pre-service biology teachers	positive effects of variables *number of natural sciences* (1 or 2) and *academic level* (Bachelor or Master) on test scores	Mathesius et al. 2016

Source of validity evidence	Investigation of …	Instrument and sample	Results	Reference
Relations to other variables	… groups with hypothesized mean equivalence	63 single best answer items, $N = 184$	highly similar ability distributions but no significant equivalence effect	Hartmann, Ziegler et al. 2019
	… correlations with I-S-T-screening and the complex problem-solving micro world	123 single best answer items, I-S-T-screening, 12 GL-micro world problems, $N = 232$	24 % of variance are explained by the investigated variables	Mathesius et al. 2019
	… correlation with a conceptually related construct	15 single best answer biology items, 12 multiple-choice biology items, I-S-T-screening, $N = 65$	substantial correlation between the two instruments ($r = .49$) that remains stable if controlled for general cognitive ability ($r_{partial} = .48$)	Hartmann, Krüger et al. 2019

4.6 Competence Development

A general indicator for the sensitivity of the test scores on learning opportunities can be inferred from the longitudinal data we collected over the time span of pre-service science teachers' academic training at Freie Universität Berlin and Humboldt-Universität zu Berlin. The participants of the longitudinal *Ko-WADiS* study had to process the test at four times: at the beginning of their academic studies, in their fourth undergraduate semester, at the beginning of their postgraduate studies, and at the fourth postgraduate semester (which is usually the semester in which they graduate as a Master of Education). Data collection took place in regular academic seminars and lectures.

Utilizing a cohort-sequential longitudinal design, three complete cohorts with a total of 644 students were tested. Due to the balanced-incomplete block design, we utilized IRT models to estimate our students' competencies. WLE were used as measures of person ability. As students at the participating universities can choose freely in which semester they apply for certain courses, additional data collection

took place unsystematically at times different from the first and fourth semesters. Missing data in the longitudinal panel was imputed using the MICE procedure (multiple imputation with chained equations; van Buuren and Groothuis-Outshoorn 2011). The results show a moderate increase of competencies over time (Figure 1).

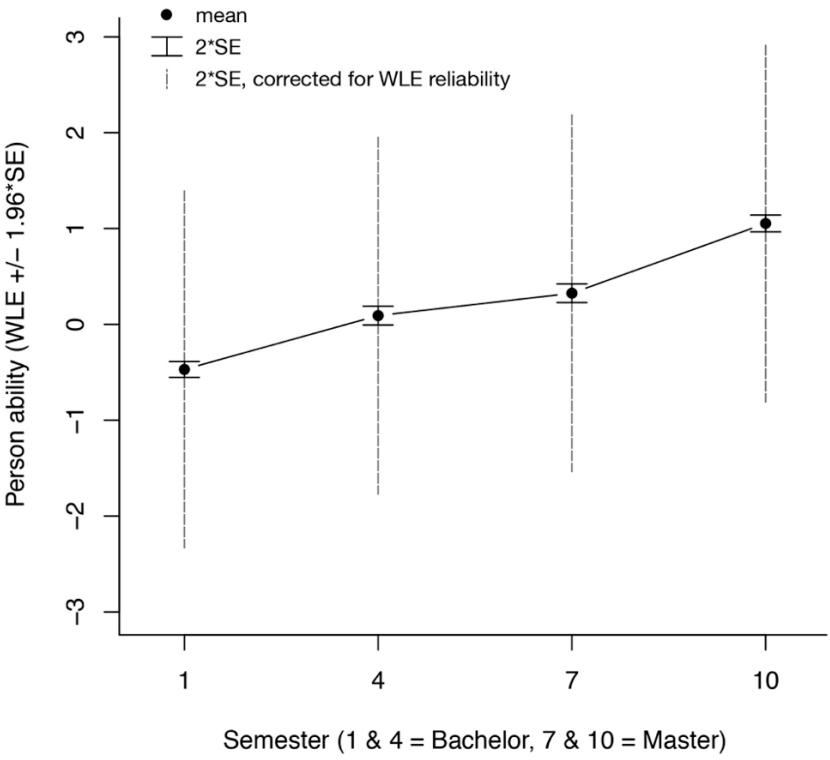

Figure 1 Development of 644 pre-service science teachers' scientific-reasoning skills during academic education. Means of weighted likelihood estimates (WLE), twofold standard errors, and twofold standard errors corrected for WLE reliability

Given that the competencies in question are part of the students' academic training, this increase is in accordance with our test score interpretation. However, the validity evidence that can be derived from this finding is limited, as other competencies increase during academic education as well.

5 Reliability

During the longitudinal *Ko-WADiS* assessment and the investigation of validity in the *ValiDiS* project, several person and item samples and subsamples were tested. Therefore, different measures of reliability were investigated. The reliability measures vary depending on the particular sample considered, but overall are satisfactory:

The rating scale used to investigate expert judgments had a *Cronbach's* α of .63 (six items, $N = 21$ experts). Dimensionality analyses were based on data from 3 010 pre-service science teachers and science students, utilizing a one-parametric logistic model with latent ability estimates. The according Expected-A-Posteriori/Plausible Value (EAP/PV) reliability is 0.47.

The test scores' instructional sensitivity to learning progress in regular lectures was investigated by comparing Weighted Likelihood Estimates (WLE) as measures of the 59 participants' individual abilities. The corresponding WLE reliability of the long version of the test (123 items) is 0.40, and the test-retest reliability is .60. Instructional sensitivity was further tested in an experimental intervention with 87 participants. The nine-item short version of the test has a *Cronbach's* α of .44 and a test-retest reliability of .48.

The test's sensitivity to known-groups differences was modelled as latent regression with latent estimates as measures for group means and variances. The corresponding EAP/PV reliability is 0.54 ($N = 2\,247$; Hartmann et al. 2015b). In a subsample of pre-service biology teachers, the EAP/PV reliability of the test was 0.66 ($N = 626$; Mathesius et al. 2016).

Hypotheses of group equivalence were tested by comparing WLE distributions in two matched samples of 131 pre-service science teachers and 131 psychology students. The estimation was based on the optimized 63-item version of our instrument, with a WLE reliability of 0.59.

The short version of our instrument used to investigate the convergent validity (15 items) has a *Cronbach's* α of .61 ($N = 65$ pre-service biology teachers). The 21 item biology test booklet has a *Cronbach's* α of .60.

Investigating the empirical relationship between the *Ko-WADiS* test (120 items) with the I-S-T 2000 R and Genetics Lab test (Mathesius et al. 2019), the corresponding EAP/PV reliability was 0.55 ($N = 232$). Finally, the instrument used to investigate competence development (123 items, $N = 644$) has a WLE reliability of 0.50.

Overall, the reliabilities found in our studies are comparable to the values of other projects which reported reliabilities for scientific-reasoning tests between 0.23 and 0.66 (e.g., Mannel 2011; Neumann 2011; Wellnitz 2012).

6 Discussion and Implications

The *Ko-WADiS* competence test provides test takers, academic teachers and educational researchers with an instrument to measure competencies and their development reliably, validly and economically. Multiple evidences of validity support the interpretation of the test scores as a measure of scientific reasoning competencies. Longitudinal results show an increase in competence during academic education.

The intended use of the *Ko-WADiS* competence test is to provide an empirical basis to describe pre-service science teachers' scientific-reasoning competencies and the development of these competencies. The test was specifically designed for large-sample scenarios, such as monitoring studies. The use of the instrument for individual diagnoses is not intended and therefore has not been investigated. However, short versions of the test were used for the assessment of validity in relatively small samples (Hartmann et al. 2019).

A future perspective on test use might be the question how to further develop teaching and learning scientific reasoning competencies. We assume that, starting from our rather broad construct, learning opportunities focusing on specific aspects of scientific reasoning could be evaluated with short versions of the test. For example, the effects of a seminar about scientific modeling on students' modelling competencies could be investigated by utilizing a short version of our instrument that consists of test items from the modelling subscale. However, in such a scenario, the reliability and validity must be investigated again before inferences are drawn from the results.

In terms of dissemination such short-test might be used – eventually adapted – for the purpose of experimental intervention studies with a specific theoretical background and research questions. Such an approach would give additional insight into test characteristics, but at the same time would help to develop teaching and learning. Digitalization of the test – eventually in an adaptive way – might help dissemination (Brüggeman and Nordmeier 2018). The dissemination of the test use or the use of short-tests means to evaluate possible transfer of scientific reasoning competencies.

The effectiveness of seminars fostering scientific reasoning competencies can also be investigated in interventions using the *Ko-WADiS* test (Mathesius et al. in preparation). The *Ko-WADiS* test has not been developed for individual diagnosis. Nevertheless, a computer-aided adaptive test with three test blocks of five items each is in preparation to enable individual diagnose. The goal is to make the measurement of scientific reasoning competencies more economical while maintaining the same reliability and validity, and to enable individual diagnostics (Brüggemann and Nordmeier 2018).

Limitations arise from the rather low reliability of the test scores. Though not unusual for tests of this kind, the mediocre consistency measures indicate a notable amount of standard error, which significantly affects the outcome of inferential analyses. Paired with the relatively small effect sizes we found in almost all scenarios, the potential of using the test in small samples is limited.

The discussion of small effect sizes brings up two different strands of argumentation. Either, the assumption of unidimensionality of the scales goes along with low reliabilities that are explained by construct-irrelevant variance. Following Cronbach (1951), the reliability should be calculated separately for the items relating to different sub-dimensions. In this case, we would face the problem of construct-underrepresentation as we wouldn't have enough items in each test booklet. Therefore, a possible interpretation for issues of reliability of a research instrument could be that students' responses are highly situated and contextualized (Leach et al. 2000). Adams and Wieman (2011) pick up on this point by arguing that a high correlation between tasks means that the tasks are repetitive. The observation that in the preceding validation analyses students understood the items as intended and gave reasonable explanations for their responses (Mathesius et al. 2018) could indicate that the low reliability is a consequence of the students' diverse understanding across the sub-dimensions. However, the dimensionality analyses as well as correlations between items and between sub-competences do not foster this interpretation (Hartmann et al. 2015a).

With regard to international analyses, the 21 biology items of the *Ko-WADiS* instrument were translated into English, Spanish and Greek (TRAPD: Translation, Review, Adjudication, Pretest, Documentation, Harkness et al. 2010) and assessed in Australia (Krell et al. 2018), Chile (Krell et al., in preparation), Canada[3] and Cyprus. The translation into French is currently taking place. As far as investigated, only a few, already revised items in other languages led to moderate DIFs (Krell et al. 2018).

References

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education, 33*(9), 1289–1312.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education [AERA, APA & NCME] (2014). *Standards*

3 Ethics committees in Germany and Canada have approved the questionnaire for use.

for educational and psychological testing. Washington, DC: American Educational Research Association.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie, 223*(1), 3–13.

Brüggemann, V., & Nordmeier, V. (2018). Naturwissenschaftliches Denken im Lehramtsstudium- Computeradaptive Leistungsmessung. In C. Maurer (Ed.), Qualitätsvoller Chemie- und Physikunterricht – normative und empirische Dimensionen. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung in Regensburg 2017 (pp. 915–918). Regensburg: Universität Regensburg.

Van Buuren, S., & Groothuis-Outshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(3), 28–45.

Giere, R. N., Bickle, J., & Mauldin, R. F. (2006). *Understanding scientific reasoning*. Independence, KY: Wadsworth/Cengage Learning.

Gilbert, J. K., & Justi, R. (2016). *Modelling-based teaching in science education* (Vol. 9). Switzerland: Springer.

Van Gog, T., Paas, F., van Merriënboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: cued retrospective reporting versus concurrent and retrospective reporting. *Journal of experimental psychology, 11*(4), 237–244.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph, 3*, 125-156.

Greiff, S., & Fischer, A. (2013). Der Nutzen einer komplexen Problemlösekompetenz. *Zeitschrift für Pädagogische Psychologie, 27*(1-2), 27–39.

Großschedl, J., Welter, V., & Harms, U. (2018). A new instrument for measuring pre-service biology teachers' pedagogical content knowledge: The PCKIBI. *Journal of Research in Science Teaching*. Advance online publication. Doi:10.1002/tea.21482

Gruber, H. (2010). Expertise. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (pp. 183–189). Weinheim: Beltz.

Gut-Glanzmann, C., & Mayer, J. (2018). Experimentelle Kompetenz. In D. Krüger, I. Parchmann & H. Schecker (Eds.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (pp. 121–140). Berlin: Springer.

Harkness, J.A., Braun, M., Ewards, B., Johnson, T.P., Lyberg, L., Mohler, P.Ph., PenneIl, B.-E., & Smith T.W. (Eds.). (2010). *Survey methods in multinational, multiregional, and multicultural contexts*. New Jersey: John Wiley & Sons.

Hartmann, S., Mathesius, S., Stiller, J., Straube, P., Krüger, D., & Upmeier zu Belzen, A. (2015a). Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung als Teil des Professionswissens zukünftiger Lehrkräfte. In B. Koch-Priewe, A. Köker, J. Siefried &

E. Wuttke (Eds.), *Kompetenzerwerb an Hochschulen: Modellierung und Messung* (pp. 39–58). Kempten: Klinkhardt.

Hartmann, S., Upmeier zu Belzen, A., Krüger, D., & Pant, H. A. (2015b). Scientific reasoning in higher education: Constructing and evaluating the criterion-related validity of an assessment of preservice science teachers' competencies. *Zeitschrift für Psychologie, 223*, 47–53.

Hartmann, S., Krüger, D., & Upmeier zu Belzen, A. (2019a). Investigating the validity and reliability of a scientific reasoning test for pre-service teachers. Vortrag auf der ESERA September 2019, Bologna.

Hartmann, S., Ziegler, M., Krüger, D., & Upmeier zu Belzen, A. (2019b). "Equivalent-groups validation"? Practical application and critical examination of a known-groups approach to investigate the criterion-related validity of test score interpretations.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*, 448–457.

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods, 6*, 81–90. Doi:10.22237/jmasm/1177992480

Klahr, D. (2000). Exploring science. *The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

KMK (Ed.). (2014). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung*. Berlin, Germany: author. Retrieved from http://www.akkreditierungsrat.de/fileadmin/Seiteninhalte/KMK/Vorgaben/KMK_Lehrerbildung_inhaltliche_Anforderungen_aktuell.pdf

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie, 216*, 61–73.

Krell, M., Redman, C., Mathesius, S., Krüger, D., & van Driel, J. (2018). Assessing pre-service science teachers' scientific reasoning competencies. *Research in Science Education*, 1–25.

Krell, M., Mathesius, S., van Driel, J., Vergara, C., & Krüger, D. (in preparation). Assessing scientific reasoning competencies of pre-service science teachers: Applying the TRAPD approach to translate a German multiple choice questionnaire into English and Spanish. *International Journal of Science Education*.

Krüger, D., Kauertz, A., & Upmeier zu Belzen, A. (2018). Modelle und das Modellieren in den Naturwissenschaften. In D. Krüger, I. Parchmann & H. Schecker (Eds.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (pp. 141–157). Berlin: Springer.

Lawson, A.E., Clark, B., Cramer- Meldrum, E., Falconer, K.A., Sequist, J. M., & Kwon, Y.-J. (2000). Development of scientific reasoning in college biology: Do two levels of general Hypothesis-testing skills exist? *Journal of Research in Science Teaching, 37*, 81–101.

Leach, J., Millar, R., Ryder, J., & Séré, M.-G. (2000). Epistemological understanding in science learning: the consistency of representations across contexts. *Learning and Instruction, 10*(6), 497–527.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Göttingen: Hogrefe.

Mannel, S. (2011). *Assessing scientific inquiry. Development and evaluation of a test for the low-performing stage*. Berlin: Logos.

Mathesius, S., Upmeier zu Belzen, A., & Krüger, D. (2014). Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung. *Erkenntnisweg Biologiedidaktik, 13*, 73–88.

Mathesius, S., Hartmann, S., Upmeier zu Belzen, A., & Krüger, D. (2016). Scientific reasoning as an aspect of pre-service biology teacher education: Assessing competencies using a paper-pencil test. In T. Tal & A. Yarden (Eds.), *The future of biology education research* (pp. 93–110). Haifa, Israel: The Technion, Israel Institute of Technology/The Weizmann Institute of Science.

Mathesius, S., Upmeier zu Belzen, A., & Krüger, D. (2018). Eyetracking als Methode zur Untersuchung von Lösungsprozessen bei Multiple-Choice-Aufgaben zum wissenschaftlichen Denken. In M. Hammann & M. Lindner (Eds.), *Lehr- und Lernforschung in der Biologiedidaktik: Band 8* (pp. 225–244). Innsbruck: Studienverlag.

Mathesius, S., Krell, M., Upmeier zu Belzen, A., & Krüger, D. (2019). Überprüfung eines Tests zum wissenschaftlichen Denken unter Berücksichtigung des Validitätskriteriums relations-to-other-variables. *Zeitschrift für Pädagogik, 65*(4), 492–510.

Mathesius, S., Bruckermann, T., Schlüter, K., & Krüger, D. (in preparation). Assessing pre-service science teachers' scientific reasoning competencies: Using known-groups as source of validity evidence for a scientific reasoning test.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In: D. Krüger & H. Vogt (Eds.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (pp. 177–186). Berlin Heidelberg: Springer.

Neumann, I. (2011). *Beyond physics content knowledge. Modeling competence regarding nature of science inquiry and nature of scientific knowledge*. Berlin: Logos.

NGSS Lead States (Ed.). (2013). *Next generation science standards: for states, by states*. Washington, DC: The National Academies Press.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265–279.

Osborne, J. (2018). Styles of scientific reasoning: What can we learn from looking at the product, Not the process, of scientific reasoning? In F. Fischer, C. A. Chinn, K. Engelmann & J. Osborne (Eds.), *Scientific reasoning and argumentation* (pp. 162–186). New York: Taylor & Francis.

Popper, K. R. (2004). *Unended quest: An intellectual autobiography*. London: Routledge.

Priemer, B., Eilerts, K., Filler, A., Pinkwart, N., Rösken-Winter, B., Tiemann, R., & Upmeier zu Belzen, A. (2019). A framework to foster scientific problem-solving in STEM and computing education. *Research in Science & Technological Education, 3*(2), 1–26.

Roberts, R., & Gott, R. (2003). Assessment of biology investigations. *Journal of Biological Education, 37*(3), 114–121.

Rönnebeck, S., Schöps, K., Prenzel, M., Mildner, D., & Hochweber, J. (2010). Naturwissenschaftliche Kompetenz von PISA 2006 bis PISA 2009. In: E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Eds.), *PISA 2009 Bilanz nach einem Jahrzehnt* (pp. 177–198). Münster: Waxmann.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*, 657–680.

Sonnleitner, P., Keller, U., Romain, M., & Brunner, M. (2013). Students' Complex Problem-solving Abilities. *Intelligence*, 41(5), 289–305.

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assessment and Evaluation in Higher Education, 41*(5), 721–732.

Straube, P. (2016). *Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-)Studierenden im Fach Physik*. Berlin: Logos-Verlag.

Trilling, B., & Fadel, C. (2009). *Twenty-first century skills. Learning for life in our times*. San Francisco: Jossey-Bass.

Upmeier zu Belzen, A., & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 41–57.

Wellnitz, N. (2012). *Kompetenzstruktur und -niveaus von Methoden der naturwissenschaftlichen Erkenntnisgewinnung*. Berlin: Logos.

Wellnitz, N., & Mayer, J. (2013). Erkenntnismethoden in der Biologie – Entwicklung und Evaluation eines Kompetenzmodells. *Zeitschrift für Didaktik der Naturwissenschaften, 19*, 315–345.

White, P. (2017). *Developing research questions*. London: Palgrave Macmillan.