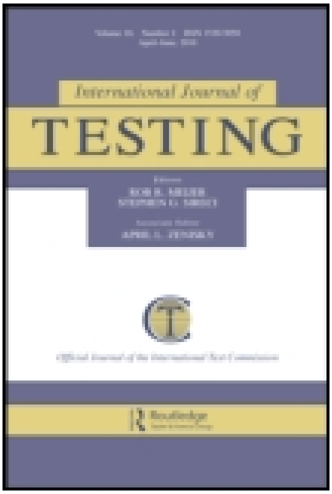

International Journal of Testing
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hijt20

Validating the Interpretations of PISA and TIMSS Tasks: A Rating Study

Heiner Rindermann[a] & Antonia E. E. Baumeister[a]
[a] Department of Psychology, Technische Universität Chemnitz, Germany

Published online: 22 Dec 2014.

To cite this article: Heiner Rindermann & Antonia E. E. Baumeister (2015) Validating the Interpretations of PISA and TIMSS Tasks: A Rating Study, International Journal of Testing, 15:1, 1-22, DOI: 10.1080/15305058.2014.966911

To link to this article: http://dx.doi.org/10.1080/15305058.2014.966911

International Journal of Testing, 15: 1–22, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2014.966911

Validating the Interpretations of PISA and TIMSS Tasks: A Rating Study

Heiner Rindermann and Antonia E. E. Baumeister
Department of Psychology, Technische Universität Chemnitz, Germany

Scholastic tests regard cognitive abilities to be domain-specific competences. However, high correlations between competences indicate either high task similarity or a dependence on common factors. The present rating study examined the validity of 12 Programme for International Student Assessment (PISA) and Third or Trends in International Mathematics and Science Study (TIMSS) tasks. Two tasks per competence (reading, mathematics, science, problem solving) from PISA and TIMSS were assessed by 34 teachers and 33 psychology students on 11 scales: difficulty, curriculum reference, knowledge versus thinking, reading competence, verbal ability, math competence, science competence, problem solving, reasoning, general knowledge, and intelligence. Intraclass correlation between two randomly chosen raters was $r_{\text{ic}} = .59$. None of the tasks represented the intended target competence concisely. In five PISA tasks, competences other than those intended were seen as being more relevant. TIMSS tasks were seen as more curriculum-related and requiring more school knowledge than PISA tasks. For solving PISA tasks, thinking/reasoning ability and general intelligence were rated as being more important ($d = 0.36$). Only small differences were found between students' and teachers' ratings.

Keywords: *intelligence, PISA, student assessment tests, TIMSS, validity*

INTRODUCTION

In the context of the international student assessment studies (SAS) Third or Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA), and Progress in International Reading Literacy Study (PIRLS), competence models have been developed to create

We want to express our gratitude to Nadja Kozel, Tamara Steiner, and Sigrid Wimmer for collecting the data.

Correspondence should be sent to Prof. Dr. Heiner Rindermann, Technische Universität Chemnitz, Department of Psychology, Wilhelm-Raabe-Str. 43, D-09107 Chemnitz, Germany. E-mail: heiner.rindermann@psychologie.tu-chemnitz.de

competence scales (e.g., Hartig & Klieme, 2006; Organisation for Economic Co-operation and Development [OECD], 2003). These models distinguish between reading, mathematics, and science literacy (problem solving was added in 2003 and 2012). Based on content-related and cognitive psychological assumptions, further distinctions have been made within the different competences. Additionally, competence levels (according to task difficulty) were defined and different response formats (open vs. closed) were used. However, researchers from different paradigms, for example, domain-specific educational research (e.g., mathematics education) or psychometric intelligence research, have raised concerns regarding the underlying competence models of SAS scales. Points of criticism were, for example, that important content was either not assessed or assessed incorrectly, competence levels were not modeled accurately, and model assumptions were not testable (e.g., Freudenthal, 1975; Hopmann, Brinek, & Retzl, 2007; Jahnke & Meyerhöfer, 2007; Wang, 1998).

In addition, several authors, including authors involved in student assessment studies, have found high correlations across different competences (manifest $r \geq .50$, latent $r_l \geq .80$; e.g., Brunner, 2008; OECD, 2014, p. 68). Applying a factor analysis to mathematical achievement tasks, Kobarg and Dalehefte (2012) found a stronger intelligence factor (mean loading $\lambda = .42$) than a mathematical achievement factor (mean loading $\lambda = .31$). One would expect that PISA Math correlates higher with grades in Mathematics than with grades in German but this is not the case ($r = .26$ and $r = .28$, respectively). A similar discordant pattern was found for PISA Math and grades in Science versus Mathematics or for PISA Science with grades in German (Fischbach, Keller, Preckel, & Brunner, 2013, their Table 3, p. 70). The sample correlations even were slightly higher with the "wrong" content area!

Rindermann (2006, 2007) pointed out that high correlations and strong *g*-factors explaining roughly 80% of the variance are found for individual differences, but also on the aggregate level, for example, across the different German federal states, and internationally, across different countries. This means that there is not only a robust common factor for persons (e.g., "If he or she is good in reading, then he or she is also good in mathematics"), but also for the German federal states (e.g., Bavarian pupils score higher on all competence dimensions than pupils from Bremen) and for countries (e.g., Finnish pupils score higher on all competence dimensions than Spanish pupils). Furthermore, not only are there empirical relationships between different competences within student assessment studies but also associations between these competences and other cognitive ability measures, for example, psychometric intelligence tests (similarly for other types of student assessments: Ceci, 1991; Deary, Strand, Smith, & Fernandes, 2007; Frey & Detterman, 2004; Jensen, 1989; Steinmayr & Meißner, 2013). This is not surprising: low correlations would have been implausible because various factors like the following contribute to these empirically observed relationships:

1. Genetic factors most probably have a global rather than a specific influence on cognitive competences (e.g., Haworth, Kovas, Dale, & Plomin, 2008). This includes indirect effects through selecting and shaping environments.
2. Basic cognitive competences like mental speed and working memory exert a rather global influence on complex cognitive competences (e.g., Rindermann & Neubauer, 2000; Rindermann, Michou, & Thompson, 2011).
3. Environmental factors such as physical, cultural-familial, and school environment have a rather global influence (cf. Rindermann & Heller, 2005).
4. During development, there are positive interactions between several subsystems, for example, between reasoning and knowledge, interest and competence (e.g., Maas et al., 2006; Rindermann, Flores-Mendoza, & Mansur-Alves, 2010).
5. Similar cognitive processes are required for solving tasks from different tests and scales, for example, attention control, concept formation, inductive and deductive reasoning, knowledge retrieval, and knowledge application. Intelligence and test-taking skills based on prior experience help to solve all kinds of tests (Rost & Sparfeldt, 2007).
6. In a test situation, similar personality traits are important for being successful in different tasks such as diligence and low test anxiety (e.g., Ackerman & Heggestad, 1997; Hattie, 2009).
7. Finally, with regard to content, there are similarities between items and similarities of the cognitive demands across different tasks, dimensions, and tests.

Only the last point (7) is crucial for the quality of tests, whereas all of the other aforementioned factors cannot be avoided or are only increased by content similarity (5).

If the results of scales, which are supposed to assess distinct competences, math versus reading, for example, or student achievement versus intelligence, correlate highly because their items are similar and they actually measure broader concepts than they declare to measure, this would be a problem of test construction. In contrast, if it was only intended to measure those competences that are useful for coping with the cognitive challenges of modernity, and if these cognitive challenges shared similar aspects (e.g., reading and reasoning), overlapping and therefore correlating scales would not be problematic.

At first glance, there seem to be content overlaps. For example, the PISA “Lake Chad” task or the “Flu” task (cf. OECD, 2009, pp. 17–21) cannot clearly be categorized. Both could be categorized as scientific and reading literacy tasks (actually, the developers declared them to be reading literacy tasks). Such strong overlaps regarding content are unknown in the context of intelligence tests like Cognitive Abilities Test (CogAT) (Lohman & Hagen, 2002) or Berliner Intelligenzstruktur

Test–Form 4 (Jäger, Süß, & Beauducel, 1997). Those items can clearly be assigned to the verbal, numerical, and figural scales, respectively.

In addition, PISA and TIMSS differ in how similar their tasks are to the curriculum taught at school: PISA tries to measure distinct abilities of students for coping with the cognitive challenges of modernity ("knowledge and skills essential for full participation in the knowledge society," "to meet real-life challenges"; OECD, 2009, pp. 12, 13). PISA abandons the assessment of mere academic knowledge but focuses on assessing competences, which can be acquired academically but also outside of school. PISA uses the term "literacy," which stands for competence and distinguishes three main types of competences: reading literacy, mathematical literacy, and scientific literacy, as well as problem solving. Among all constructs assessed, the problem-solving scale shows the highest similarity with intelligence (Rindermann, 2006). Further, PISA tasks comprise a lot of text. Each problem is followed by several open or closed questions (with multiple-choice answers) assessing the indicated competence in different ways. Many tasks require quite general world knowledge, which can be learned at school, but also in family or by the media (reading books, newspapers, and internet pages, watching and hearing stimulating TV and radio programs; Rindermann, 2006) instead of specific academic knowledge.

With regard to TIMSS, the assessment of the competences of numeracy and science is more closely related to the subjects taught at school. Generally, tasks of TIMSS are shorter than those of PISA. They include little or no text. Similarly to PISA, there are two answering modes: open and multiple-choice. Solving TIMSS tasks appears to depend more on knowledge retrieval. In contrast, the extensive texts and accompanying figures of PISA tasks, often including graphs with numbers, require multiple mental translations and integration resulting in a higher task complexity (Mayer, 2005; Ullrich et al., 2012).

In order to classify the structure and the requirements of the tasks, a first (unpublished) study was conducted by Brunner, Kunter, and Krauss (2005). In their study, three groups of experts (10 "mathematic educationalists," 4 "experts on psychometric research on intelligence," and 59 "experts on German mathematics curricula") rated the cognitive requirements. These ratings were compared to the results of the factor loadings of the test items on latent variables. No accordance could be found between expert ratings and the psychometric competence structure. We interpret the finding of Brunner and colleagues (2005) as evidence of the fact that tasks of the international student assessment tests show a discrepancy between concepts and empirical findings.

According to Hartig and Jude (2007, p. 22), however, only a sufficiently high content validity legitimates drawing individual and institutional consequences from applying tests. Kane (2013, p. 65) recommended developing an "interpretation/use argument" (IUA) approach for validating the "intended interpretations and uses of test scores" (Kane, 2013, p. 8). Following this, it should be possible for persons other than the test developers to classify a test according to the construct that the test developers intended to measure. One possibility for checking this is to let experts rate the content of PISA and TIMSS items. These experts should have no special prior knowledge about the items and scales, their meaning, and their theoretical explanation in order to reduce confirmatory biases. In the present study, expert ratings of teachers and university students were obtained who were generally familiar with school tasks but who, in contrast to the procedure by Neubrand, Klieme, Lüdtke, and Neubrand (2002), did not receive any instruction on the postulated competence structure models of PISA and TIMSS. Thus, these expert groups provided rather unbiased ratings on the basis of their own prior knowledge and experience regarding the content of the tasks, their specificity, reference to curriculum, difficulty, solution procedures, and demands on general intelligence.

Teachers of those subjects that are covered by PISA and TIMSS (literacy, e.g., in German, mathematics, science), have passed a university education and have collected professional experience which qualifies them to provide such ratings. Students of psychology as judges, however, are less involved in the content than teachers. Nevertheless, research on students' evaluation of instruction, for example, shows that student ratings can be regarded as both reliable and valid (Rindermann & Schofield, 2001). If students are able to rate the complex and dynamic nature of a university course relatively accurately then they should also be able to rate the content of a student assessment test. However, due to the higher expertise of teachers, deviations in the ratings cannot be excluded. The present empirical study aimed at answering these questions by investigating the validity of the interpretations of PISA and TIMSS tasks developed for 15-year-old pupils or eighth graders. In this study, the validity arguments were formulated according to the basic assessment frameworks of PISA and TIMSS (cf. Appendix A; Mullis et al., 2003, pp. 7–69; OECD, 2003, pp. 15–17). Further, general criteria of test diagnostics were used in order to examine whether the content validity argument of the PISA and TIMSS tasks is plausible. We asked teachers and university students, not the possibly biased test developers, to categorize the content and to estimate the cognitive requirements of the tasks both of which correspond to the interpretation aspect of the IUA (Kane, 2013).

Based on the aforementioned results, the following research questions were investigated: (1) Do PISA and TIMSS tasks measure what they intend to measure? We assumed that they measure more. (2) Do TIMSS and PISA tasks differ in their requirements concerning general intelligence and general knowledge? Our hypotheses were that PISA tasks require both more general intelligence and more general knowledge than TIMSS tasks. (3) Do TIMSS and PISA tasks differ with regard to the cognitive processes required to solve them? We assumed that TIMSS tasks are more likely to be solved by means of knowledge application, PISA tasks by means of reasoning/understanding. (4) Are TIMSS tasks more specific than PISA tasks with regard to their assessment of competences? We assumed that TIMSS tasks assess competences more specifically than PISA tasks. (5) Do TIMSS tasks assess more curriculum-related knowledge than PISA tasks? The hypothesis was that TIMSS tasks show a closer reference to curriculum. (6) Are there differences in difficulty between TIMSS and PISA tasks? We assumed that PISA tasks are more difficult.

METHOD

In order to test the hypotheses, eight PISA tasks and four TIMSS tasks were selected. Teachers and students were provided with the tasks and were asked to rate the tasks on several dimensions.

Participants

Ratings of 68 persons were collected. Thirty-five raters were teachers (71% females, mean age 44 years) and 33 raters were students (73% females, mean age 25 years). Due to missing data, one teacher was excluded.

The raters were chosen according to the following criteria: (1) One of the subjects of each teacher had to be either a language, mathematics, or science. The teachers had to teach students between the ages of 13 to 15 years; this corresponds to the age group that took the PISA and TIMSS tasks. Type of school ("Gymnasium, AHS" leading to university: 17 teachers, vs. "Hauptschule" and "Berufsschule" leading to vocational training: 16 teachers in total) was not important. (2) Among students, psychology students were chosen (graduate students/postgraduate master students in the second phase of their diploma studies). They already had acquired profound knowledge in diagnostics and in applying psychology tests.

Measures

The test material consisted of three parts: (1) A test booklet containing the tasks to be rated (cf. Table 1), (2) a paper sheet containing the definitions of the rating scales (cf. Appendix A), and (3) a booklet containing the rating scales (cf. Appendix B).

The 12 items of the test booklet (see Table 1) represent a selection of PISA (four competences, eight items) and TIMSS (two competences, four items) tasks. For each competence, two items were included, which ought to be as representative as possible and which ought to have a medium level of complexity. We regard those tasks that were chosen for publication by the PISA and TIMSS test authors themselves as good examples of their measurement approach. Within the published tasks we selected those including various answer formats (i.e., closed as well as open questions).

TABLE 1
Competence Ratings as Raw Scale Values

Test	PISA								TIMSS			
Scale Competence	Reading		Math		Science		Probl. Solv.		Math		Science	
Item	Lake Chad	Graffiti	Farms	Apples	Cloning	Ozone	Cinema Outing	Library System	Area Shading	Marbles	Plant Cell	Hibernation
Number	4	9	7	10	2	12	1	11	3	6	5	8
Difficulty	—	—	5.42	4.71	4.46	5.71	—	—	3.74	3.64	4.65	1.93
			(1.97)	(1.86)	(1.97)	(1.71)			(2.06)	(2.20)	(1.96)	(1.47)
Reading	*6.10*	*7.30*	4.69	4.63	6.91	7.16	6.00	6.24	1.46	4.27	3.78	4.15
competence	(1.82)	(1.22)	(2.20)	(2.17)	(1.32)	(1.19)	(2.09)	(1.98)	(1.68)	(2.37)	(2.18)	(2.21)
Verbal	*4.43*	*7.22*	3.10	3.42	5.64	6.75	4.40	5.10	1.07	2.76	3.63	3.16
competence	(2.36)	(1.20)	(2.35)	(2.37)	(2.00)	(1.67)	(2.68)	(2.28)	(1.63)	(2.20)	(2.33)	(2.23)
Math	3.27	0.21	*7.60*	*7.27*	1.03	1.49	3.40	3.55	*7.12*	*7.19*	0.33	0.37
competence	(2.33)	(0.66)	(0.87)	(1.20)	(1.50)	(1.75)	(2.24)	(2.53)	(1.50)	(1.36)	(0.81)	(1.09)
Science	4.51	0.42	2.27	2.25	*5.66*	*6.57*	0.76	1.06	1.03	1.37	*7.01*	*5.96*
competence	(2.45)	(1.02)	(2.51)	(2.50)	(1.76)	(1.59)	(1.40)	(1.84)	(1.57)	(2.02)	(1.63)	(2.09)

(Continued on next page)

TABLE 1
Competence Ratings as Raw Scale Values *(Continued)*

Test	PISA								TIMSS			
Scale Competence	Reading		Math		Science		Probl. Solv.		Math		Science	
Item	Lake Chad	Graffiti	Farms	Apples	Cloning	Ozone	Cinema Outing	Library System	Area Shading	Marbles	Plant Cell	Hibernation
Problem-solving ability	5.04 (2.29)	2.69 (2.26)	5.63 (2.19)	5.91 (2.03)	3.70 (2.28)	4.78 (2.24)	*6.45* (1.60)	*6.12* (1.87)	4.99 (2.52)	6.04 (2.05)	1.09 (1.57)	1.36 (2.03)
Reasoning	6.21 (2.03)	3.91 (2.45)	5.65 (2.06)	6.55 (1.60)	5.39 (2.39)	5.28 (2.26)	6.86 (1.35)	6.97 (1.50)	4.10 (2.86)	6.49 (1.81)	1.58 (2.04)	3.04 (2.46)
General knowledge	4.06 (2.28)	4.39 (2.41)	3.66 (2.24)	3.63 (2.35)	4.90 (1.63)	6.00 (1.33)	2.90 (2.22)	3.90 (2.36)	3.04 (2.33)	2.88 (2.27)	5.50 (2.03)	5.94 (1.59)
Intelligence	5.43 (1.79)	5.15 (1.82)	5.18 (1.66)	5.58 (1.58)	4.99 (1.96)	5.93 (1.34)	5.27 (1.86)	5.70 (1.82)	4.85 (1.79)	5.24 (1.83)	4.00 (2.39)	4.67 (2.14)
Reference to curriculum	—	—	6.82 (1.45)	5.91 (1.91)	4.94 (2.10)	4.82 (2.39)	—	—	6.37 (1.82)	6.48 (1.99)	6.40 (2.12)	5.79 (2.03)
Knowledge-Thinking	—	—	0.48 (2.60)	1.93 (1.87)	1.73 (2.00)	0.55 (2.32)	—	—	1.18 (2.40)	2.41 (1.86)	−3.45 (0.86)	−0.97 (2.36)

Note. Competence: classification according to Programme for International Student Assessment (PISA) and Third or Trends in International Mathematics and Science Study (TIMSS); Number: order of task in this study; from Difficulty to Reference to curriculum: rating scale from 0 to 8 (midpoint 4); Knowledge-Thinking: rating scale from −4 (knowledge) to +4 (thinking; midpoint 0); TIMSS usually does not name its tasks; upper row: means (and standard deviations in parentheses); in *italics*: scale corresponding to *target literacy*; underlined: highest rating value of an item. A result being intended by the test and its theory is in *italics* and underlined.

The order of the items in the test booklet was random. The first page of the booklet contained the instruction and a note concerning confidential treatment of the data. In addition, questions regarding sociodemographic data of the raters followed. These included age, gender/sex, and profession (teacher vs. psychology student). Further, teachers were asked to indicate the type of their school.

The raters were asked to read student assessment tasks and to assess them on several dimensions. They did not have to solve the PISA and TIMSS tasks, but they should rate their content and the cognitive competences and processes required for solving the tasks on eight 9-point Likert scales (from 0 = "not at all, very low" to 8 = "completely, very high"): reading competence, verbal competence, mathematics competence, science competence, problem-solving ability, reasoning, general knowledge, and general intelligence (cf. Appendix B). Additionally, since it was possible to compare math and science items across PISA and TIMSS, those items were rated regarding task difficulty, relatedness to curriculum, and on a knowledge-thinking-continuum. Only for the last dimension, a scale ranging from –4 (knowledge) to +4 (thinking), with 0 standing for the midpoint of the scale, was used. Raters were given a one-page description of the rating scales (see Appendix A). Note that these descriptions correspond to the basic assessment frameworks of TIMSS and PISA (cf. Mullis et al., 2003, pp. 7–69; OECD, 2003, pp. 15–17). That is, the descriptions of the rating scales provided general information about the content that should be covered by the tasks according to PISA and TIMSS. The source of the tasks (PISA or TIMSS, intended scale) was not indicated for any of the tasks.

Procedure

The teachers were recruited at schools (via headmaster of the school). The university students were recruited via notice board at the institute of psychology of an Austrian university. After welcoming the participants, they received the test booklet, the booklet with the answering sheets (cf. Appendix B), and the paper sheet with the scale definitions (cf. Appendix A). The participants were asked to read the instruction first. After this, tentative questions of the participants concerning the procedure were answered. Then the participants started to judge the tasks independently. Answering the rating scales lasted between 20 and 60 minutes. At the end of this rating study, the participants were thanked and informed about the rationale and the research questions.

Analysis

Mean comparisons of judges' ratings of each PISA versus TIMSS task using Cohen's *d* are presented ($[M_A\text{-}M_B]/SD$, whereby *SD* is calculated as the mean of SD_A and SD_B). By convention, *d* effect sizes of 0.2, 0.5, and 0.8 are interpreted as small, medium, and large, respectively (Cohen, 1988). Usually, we do not recommend significance testing because it is not regarded to be useful for detecting generalizable results (e.g., Gigerenzer, 2004; Hunter, 1997). More important is to proof the stability of results across different task and rater samples and approaches (see Discussion). However, at two reviewers' request we added information about statistical significance for all *d*-values (paired two-sample *t*-tests).

Individual ratings were weighted according to the general rating tendency of a rater: First, ratings on reading competence and verbal competence were averaged (because both cover related constructs) as were ratings on reasoning and general intelligence (because reasoning is a component of general intelligence). Six scales remained (verbal, mathematics, science, problem solving, intelligence, general knowledge). Second, the ratings on these six scales were summed up. The individual scale rating was divided by the individual sum of the six scales. The result was multiplied by 100. The final value (a percentage value between 0 and 100) reflects the weight that a dimension had for the individual rater in relation to all further dimensions rated. Thus, general rating tendencies of raters to give "low" or "high" scores across all dimensions which could have biased the results are corrected statistically.

RESULTS

Inter-Rater Reliability

Inter-rater reliability was calculated by correlating the 67 raters across all item and answer scales ($N = 120$). Individual inter-rater agreement was calculated by using intraclass correlation (mean differences between raters were considered and, thus, decreased correlations) (Shrout, 1993). This single rater agreement between two randomly chosen raters was $r_{\text{ic}} = .59$. This indicates a relatively high individual agreement.

For students only ($N = 33$) it was $r_{\text{ic}} = .58$, for teachers ($N = 34$) the agreement was slightly higher with $r_{\text{ic}} = .60$. The ratings of TIMSS tasks achieved higher agreement than the ratings of PISA tasks ($r_{\text{ic-TIMSS}} = .66$ vs. $r_{\text{ic-PISA}} = .53$). At the level of rating scales, a higher agreement emerged regarding the curriculum orientated scales from reading to science including difficulty than regarding the more general scales (the first five scales averaged: $r_{\text{ic}} = .64$, $\alpha = .99$; the last six scales averaged: $r_{\text{ic}} = .32$, $\alpha = .94$).[1] Content and difficulty seem to be easier to estimate than more abstract concepts such as knowledge and thinking. However,

[1] Difficulty ($r_{\text{ic}} = .61$, $\alpha = .99$), Reading competence ($r_{\text{ic}} = .58$, $\alpha = .99$), Verbal competence ($r_{\text{ic}} = .53$, $\alpha = .99$), Math competence ($r_{\text{ic}} = .80$, $\alpha = 1$), Science competence ($r_{\text{ic}} = .67$, $\alpha = .99$), Problem-solving ability ($r_{\text{ic}} = .53$, $\alpha = .99$), Reasoning ($r_{\text{ic}} = .48$, $\alpha = .98$), General knowledge ($r_{\text{ic}} = .28$, $\alpha = .96$), Intelligence ($r_{\text{ic}} = .06$, $\alpha = .80$), Reference to curriculum ($r_{\text{ic}} = .14$, $\alpha = .91$), and

since we focus on the reliability of the results of all raters—the aggregated means or the "average rater"—with reliabilities at $\alpha = .99$ and $\alpha = .94$ (range $\alpha = .80$ to 1), the data basis is at least satisfactory and on average excellent.

Task Content Specificity?

For analyzing whether PISA and TIMSS tasks assess the declared target literacy specifically, the results of those scales that intended to measure a specific target competence were compared to the results of those scales that intended to measure other target competences (cf. Table 1, *italicized* are the means of the *target scales*, underlined are the highest rating values). Consider that all items measure both their target competence and other competences. This is not surprising. Of course, for reading texts, for example, reasoning is also required. Further, it is comprehensible that at least some of the many science tasks require dealing with numbers. The crucial question rather is how large the involvement of other competences is.

In five of eight PISA tasks (62.50% of the tasks), other competences were rated to be more relevant than the declared target ability: In one reading task (Lake Chad), reasoning instead of reading competence; in science (Cloning and Ozone), reading literacy instead of science competence; in problem solving (Cinema Outing and Library System), reasoning instead of problem solving.

Further, for analyzing the relative (percentage) scores of ratings, another method was chosen (cf. Table 2). The interpretation of the task content was regarded as being sufficiently valid if the following two criteria were met:

a. The specific target ability should constitute ≥30% of the average individual rating sum across the rating scales.

b. Each further competence should constitute <20% of the individual rating sum; thus, the declared target ability should reach the highest score and the others should be considerably smaller.

The results (Table 2) show that none of the 12 tasks fulfilled both criteria to ensure that the interpretation of the task content was sufficiently valid. Nine tasks (seven out of which were from PISA) did not fulfill criterion a (target ability ≥ 30%), and none of the tasks fulfilled criterion b (each further competence < 20%). If criterion b is relaxed to < 25%, five tasks fulfill this criterion (mathematics: PISA/Farms; PISA/Apples; TIMSS/Area Shading; science: PISA/Cloning; PISA/Ozone). Regarding the question of how many tasks of the three main target

Knowledge-Thinking ($r_{\text{ic}} = .43$, $\alpha = .98$). "r_{ic}" stands for the average agreement among two raters and is an indicator of individual rater reliability, "Cronbach-α" stands for the reliability of the average of 69 raters. Due to the large rater sample of 67 raters, the reliability of the average is quite high (cf. Shrout, 1993; Rindermann & Schofield, 2001).

TABLE 2
Competence Ratings as Percentage Values

Test	PISA								TIMSS			
Scale Competence	Reading		Math		Science		Probl. Solv.		Math		Science	
Item	Lake Chad	Graffiti	Farms	Apples	Cloning	Ozone	Cinema Outing	Library System	Area Shading	Marbles	Plant Cell	Hibernation
Reading/ verbal competence	*22*	*40*	15	15	26	23	23	23	7	15	20	18
	(10)	(14)	(6)	(6)	(7)	(4)	(7)	(7)	(7)	(7)	(11)	(9)
Math competence	10	1	*27*	*25*	3	4	12	13	*33*	*26*	1	1
	(6)	(2)	(11)	(10)	(5)	(5)	(7)	(13)	(14)	(9)	(3)	(4)
Science competence	15	2	7	7	*20*	*20*	3	3	4	4	*33*	*27*
	(7)	(4)	(7)	(7)	(5)	(5)	(4)	(5)	(5)	(6)	(11)	(11)
Problem-solving ability	17	12	19	19	12	14	*25*	*22*	20	21	4	5
	(7)	(9)	(7)	(5)	(7)	(6)	(6)	(6)	(10)	(6)	(6)	(7)
Reasoning/ intelligence	24	27	21	23	21	20	27	27	24	25	17	22
	(9)	(7)	(5)	(5)	(5)	(4)	(6)	(10)	(6)	(6)	(9)	(9)
General knowledge	13	20	12	11	18	19	10	13	12	9	25	27
	(6)	(10)	(7)	(7)	(5)	(4)	(8)	(7)	(9)	(7)	(8)	(9)

Note. Competence: classification according to Programme for International Student Assessment (PISA) and Third or Trends in International Mathematics and Science Study (TIMSS); Reading and Verbal competence combined (by taking the upper value, respectively), Reasoning and Intelligence combined (by taking the upper value, respectively); percentage values: each column adds up to 100% except for rounding errors; upper row: means (and standard deviations in parentheses); in *italics*: scale corresponding to *target ability*; underlined: highest rating value of an item. A result being intended by the test and its theory is in *italics* and underlined. Benchmark: percentage differences at 4% and larger are significant at the 1% level, percentage differences at 3% are significant at the 5% level.

competences (i.e., reading, mathematics, science) were rated according to their declared target ability in these three dimensions, four out of six PISA tasks (reading and mathematics) and all four TIMSS tasks hit the highest ratings in the domain of their declared target ability.

To sum up, the interpretation of the task content of PISA and TIMSS tasks does not seem to be very valid. Compared to other competences, they seem to assess the declared target ability only to a small extent.

PISA vs. TIMSS: General Intelligence, General Knowledge, or Specific Knowledge?

Comparing PISA and TIMSS, we assumed that intelligence is more important for solving PISA tasks than TIMSS tasks. For this purpose, the relative ratings were used (individual ratings of intelligence and knowledge weighted according to the general rating tendency of a rater). If we consider all PISA tasks (including reading and problem solving which are not covered by TIMSS), the judges rate them to require more intelligence than TIMSS tasks (mean percentage: 23.65% [$SD_\mathrm{P} = 3.72$] vs. 22.20% [$SD_\mathrm{T} = 4.24$]; 1.05% difference or $d = 0.36$, paired t-test, $p = .049$). If only the mathematics and science tasks are considered, the pattern of results turns around (21.15% [$SD_\mathrm{P} = 2.87$] vs. 22.20% [$SD_\mathrm{T} = 4.24$]; -1.05% difference or $d = -0.30, p = .065$).

Moreover, it was assumed that PISA tasks require more general knowledge than TIMSS tasks. The ratings regarding general knowledge reveal the following mean percentage scores across all selected PISA and TIMSS tasks: 14.33% [$SD_\mathrm{P} = 4.17$] vs. 18.14% [$SD_\mathrm{T} = 5.07$]; -3.81% difference or $d = -0.83$, $p < .001$; without reading and problem solving (i.e., the same scales for PISA and TIMSS): 14.71% [$SD_\mathrm{P} = 3.91$] vs. 18.14% [$SD_\mathrm{T} = 5.07$]; -3.42% difference or $d = -0.76$, $p < .001$. TIMSS tasks are clearly judged to require more general knowledge for solving the tasks.

To sum up, regarding the included reading and problem solving tasks, PISA tasks were judged to be more closely related to aspects of intelligence. TIMSS tasks were judged to require knowledge more strongly than PISA tasks.

PISA vs. TIMSS: Knowledge or Thinking?

Regarding the question of whether more knowledge or more thinking is required for solving the tasks, we used one scale with opposing poles for knowledge and thinking (only applied for competences covered both by PISA and TIMSS, i.e., mathematics and science). A scale value of 0' represents an equivalence of knowledge and thinking, a scale value of –4 represents maximum knowledge application, and a scale value of +4 maximum thinking/understanding (see Table 1).

On average, the solution strategies of PISA tasks involve more thinking than knowledge application ($M_{\text{P}} = 1.17, SD_{\text{P}} = 1.15$). In contrast, the solution strategies of TIMSS tasks involve more knowledge application than thinking ($M_{\text{T}} = -0.22$, $SD_{\text{T}} = 1.00$). The standardized difference is even more pronounced: $d = 1.29$, $p < .001$ (calculation: difference of 1.39 raw scale values divided by the averaged standard deviations). That is, the solution strategies for TIMSS tasks require more knowledge application, whereas the solution strategies for PISA tasks require more thinking and understanding.

TIMSS More Specific Than PISA?

In order to answer the question whether TIMSS tasks assess competences more specifically than PISA tasks, the percentage scores of the declared target scales were averaged for the PISA and TIMSS tasks. Regardless of whether reading and problem solving tasks are included in the PISA item pool, TIMSS tasks are not only rated to require more knowledge application and curriculum-related knowledge but also to be more specific than PISA tasks (mean percentage scores: 25.13% [$SD_{\text{P}} = 4.87$] vs. 29.57% [$SD_{\text{T}} = 8.43$]; −4.44% difference in favor of TIMSS or $d = -0.67$, $p < .001$, without reading and problem-solving tasks: 23.20% [$SD_{\text{P}} = 5.86$] vs. 29.57% [$SD_{\text{T}} = 8.43$]; −6.37% difference in favor of TIMSS or $d = -0.89$, $p < .001$). TIMSS tasks are more specific than PISA tasks.

TIMSS More Closely Referred to Curriculum?

It was postulated that TIMSS more closely assesses curriculum-related knowledge than PISA. Again, the ratings and their comparison were only conducted for competences assessed both by PISA and TIMSS, that is, mathematics and science (cf. Table 1). On a scale from 0 to 8, TIMSS tasks, on average, were rated to refer more strongly to the curriculum ($M_{\text{T}} = 6.26$, $SD_{\text{T}} = 1.30$) than PISA tasks ($M_{\text{P}} = 5.62$, $SD_{\text{P}} = 1.15$). The difference equals 0.62 scale points or $d = 0.50$, $p < .001$. TIMSS tasks more closely assess curriculum-related knowledge.

PISA Tasks More Difficult?

PISA tasks were rated to be more difficult ($M_{\text{P}} = 5.07$, $SD_{\text{P}} = 1.29$ vs. $M_{\text{T}} = 3.47$, $SD_{\text{T}} = 1.33$) than TIMSS tasks (we compared only mathematics and science). On a scale from 0 to 8, the scale difference equals to 1.60 points ($d = 1.22$, $p < .001$). PISA tasks were unequivocally rated to be more complex.

Rater Subgroups?

Correlational analyses were conducted for the rater groups (1 = psychology students, 2 = teachers; i.e., positive correlations would stand for higher teacher values) and scales. Teachers rated the relevance of education as a solution strategy to be smaller (for PISA: $r_P = -.24$; for TIMSS: $r_{T=} -.46$). The correlations with "general intelligence" range between $r = -.13$ and $r = .10$. The correlations with "knowledge vs. thinking" range between $r = -.00$ and $r = .16$. Regarding "specificity" and "reference to curriculum," the correlations turned out to be slightly negative ($r_{\text{Pall}} = -.17$ and $r_{\text{PMN}} = -.11$; for TIMSS: $r_T = -.08$; PISA: $r_P = -.04$; TIMSS: $r_T = -.18$). That is, the more curriculum-experienced teachers judged the tasks to be less educational, somewhat less specific and curriculum-related.

In addition, teachers rated the tasks to be more difficult (PISA: $r_P = .27$; TIMSS: $r_T = .12$). Here within the teachers' group, large differences were found for the two school types "Hauptschule" (vocational-technical track; $N = 14$, coded by 1) and "Gymnasium" (university-preparatory track; $N = 17$, coded by 2): Vocational-technical track teachers judged the tasks to be more difficult (PISA: $r_P = -.62$; TIMSS: $r_T = -.49$; $N = 31$). Further, the PISA tasks, which all raters generally estimated to be more difficult, were rated to be especially difficult by vocational-technical track teachers.

Most differences between the subgroups were small and unsystematic. In those cases in which the differences were systematic (i.e., stable across PISA and TIMSS) or larger, they reflected the larger expertise of teachers supporting the validity of the results (i.e., less "education," "specificity," and "reference to curriculum"), and their experience with regard to teaching different populations of pupils.

DISCUSSION

Teachers and students rated a total of 12 PISA and TIMSS tasks on 11 rating scales regarding the content that was supposed to be measured and the cognitive processes that were required. Apart from the intended content and processes (i.e., the declared literacy scales), other contents and processes were also involved in all tasks. For the majority of scales, namely five out of eight PISA tasks, the raters judged other contents and processes to be more predominant than the content targeted at by the task developers. Regarding the four TIMSS items, no such deviating categorization was found, the intended content and processes were identifiable for the raters.

If criterion a stating that the target ability should constitute at least 30% of the individual rating sum and criterion b stating that no other ability should constitute 20% or more of the individual rating sum is chosen in order to attain a sufficiently valid interpretation of the task content, then no task hit its target ability. It has to be concluded that the student assessment tasks of PISA and TIMSS assess more than only their declared domain-specific target ability ("literacy").

Reading literacy plays an important role for understanding all PISA tasks that were investigated. In contrast, reading literacy is less relevant for TIMSS tasks. The present rating study corroborates previous findings applying other methods (correlations and factor analyses using different data sets, task analyses; Rindermann, 2006, 2007) and studies conducted from other researchers (Brunner et al., 2005) showing that the validity of the interpretation of PISA and TIMSS tasks is questionable. Additionally, previous correlational and factor analytic studies showing high correlations and strong *g*-factors are supported and are explained, at least partly, by task and cognitive requirement similarity.

As expected, intelligence (reasoning, thinking, and understanding vs. knowledge) is somewhat more required for solving PISA tasks than for solving TIMSS tasks. PISA tasks generally require more intelligence than knowledge. Therefore and from a scientific perspective any avoidance of the term "intelligence" by the OECD authors is not justified (Thompson, 2013). Moreover, PISA tasks are judged to be more difficult. In contrast, knowledge and knowledge application are more important for TIMSS tasks compared to PISA tasks. Further, TIMSS tasks are seen to be more closely related to curriculum and to be more specific.

With regard to the teachers' difficulty ratings, expected school type differences between "Gymnasium" (university-preparatory track teachers) and "Hauptschule" (vocational-technical track) could be observed: Vocational-technical track teachers rated the tasks to be more complex. These findings underscore the validity of the survey. Further, the rating study revealed that the judgments of teachers and psychology students were similar.

Less reference to curriculum, thereby less reference to knowledge and larger investment of thinking (reasoning, intelligence) result in higher task difficulty. PISA tasks that require combining different sources of information present a larger cognitive challenge than routine tasks trained at school, the latter being more frequently used by TIMSS. TIMSS can be called a "school knowledge test." In contrast, PISA tasks require less knowledge and more thinking as a solution strategy. Thus, PISA tasks are testing general cognitive competence. Conventional psychometric intelligence tests use shorter items and present much less information, but the cognitive processes required for solving them and the cognitive abilities measured are much less diverging.

To conclude, PISA and TIMSS tasks do not measure the declared target ability concisely enough according to the criterion of specificity. They include a wider range of cognitive aspects, and they are highly related to the concept of intelligence. However, the tasks would not be identical to intelligence tasks if intelligence was interpreted as pure thinking ability (and not as crystallized intelligence including knowledge, according to Cattell, 1987/1971). If PISA and TIMSS had intended to assess general competence exclusively, thinking ability together with the repository and retrieval of relevant and true knowledge and its comprehensive use, the results would have supported their theoretical approach. This kind of cognitive competence is useful for dealing with the cognitive challenges of modernity. According to such a perspective, "specificity" would not be considered being crucial.

A rating study with 67 teachers and students cannot answer all concluding questions being raised. Our study was intended to be an initial attempt of a rating approach for PISA and TIMSS tasks. In future studies, first of all, other tasks than the 12 selected ones should be considered in order to check generalizability including more recently developed tasks of PISA 2009 and TIMSS 2011. PIRLS tasks could be compared with TIMSS tasks at class level 4. Other kinds of experts, for example, didactics experts, could be asked. However, the more prior knowledge such experts have about PISA, TIMSS, and PIRLS, the more their rating could be influenced by the knowledge of tasks (confirmatory bias; Kane, 2013). Replication studies with similar but not identical task and expert samples and statistical procedures should check the stability and generalizability of the results.

Another approach would be a correlational study, applying complete test booklets of diverse student assessment and intelligence tests in order to examine loading patterns and explained variances by a *g*-factor versus specific factors. If items and scales of different intelligence tests load on an "intelligence" factor more strongly than on a "student assessment factor," and items and scales of different student assessment tests more strongly on a "student assessment factor" than on an "intelligence" factor, this would support the notion of distinguishable concepts. Hierarchical models can be used to examine whether different student assessment tests (e.g., TIMSS and PISA taken together) and different intelligence tests (e.g., Raven's Matrices and CogAT taken together) form two different factors (correctly assigned to a student achievement factor vs. an intelligence factor), versus only one global factor, or mixed factors. Strong general factors and high multiple factor loadings would indicate that student assessment tests are as similar to intelligence tests than intelligence tests are similar to student assessment tests—both would assess cognitive competences (a "thinking-knowledge competence": intelligence, relevant and true knowledge, and the intelligent use of this knowledge).

It should be noted that the approach used by PIRLS, PISA, and TIMSS for estimating the abilities may increase the correlations between scales per se: The students do not receive complete test batteries but single tasks, and frequently only tasks from single scales (multimatrix sampling). For example, in PISA 2006, only 6 out of 13 test booklets contained tasks from all three literacy domains, two contained only science tasks. Plausible scale values (including those values that were not covered by the items of a single booklet) are estimated based on (1) the answers referring to the assigned literacy items, (2) the answers referring to items of the other literacy scales, and (3) the answers obtained by the questionnaire on individual and family characteristics. It is claimed that this estimation procedure would not inflate correlations. However, this claim needs to be validated by a study in which all tasks are administered, and the scale correlations are compared to the ones obtained by the multimatrix designs. Nevertheless, any methodological "weakness" would never be solely responsible for correlations among different student assessment scales and with psychometric intelligence measures: Among the aforementioned seven factors responsible for empirical correlations, only one or two are methodological; the others are common genetic, environmental, and process factors.

A correlational study by Brunner (2008) presented first evidence: The average manifest and latent correlations between PISA-Reading and PISA-Math were smaller than their average correlations with psychometric CogAT-scales (manifest $r_{\text{PR-M}} = .45$ vs. $r_{\text{P-C}} = .47$, latent $r_{\text{PR-M}} = .80$ vs. $r_{\text{P-C}} = .86$). Further, the thinking-aloud method could help revealing the cognitive processes that are required for solving the tasks (knowledge related to school instruction, general knowledge, thinking). Finally, in-depth task analyses should be used for separating the specific and general aspects, knowledge, and thinking.

REFERENCES

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*, 219–245.

Brunner, M. (2008). No *g* in education? *Learning and Individual Differences*, *18*, 152–165.

Brunner, M., Kunter, M., & Krauss, S. (2005, August). *Who can judge reasoning and knowledge demands of mathematics problems*? Paper presented at EARLI 2005, Nicosia, Cyprus.

Cattell, R. B. (1987/1971). *Intelligence: Its structure, growth and action*. Amsterdam, The Netherlands: Elsevier.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? *Developmental Psychology*, *27*, 703–722.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13–21.

Fischbach, A., Keller, U., Preckel, F., & Brunner, M. (2013). PISA proficiency scores predict educational outcomes. *Learning and Individual Differences*, *24*, 63–72.

Freudenthal, H. (1975). Pupils' achievement internationally compared. *Educational Studies in Mathematics*, *6*, 127–186.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or *g*? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science*, *15*, 373–378.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587–606.

Hartig, J., & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle [Empirical assessment of competences and psychometric competence models]. In J. Hartig & E. Klieme (Eds.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 17–36). Berlin, Germany: Bundesministerium für Bildung und Forschung.

Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik [Competence and competence assessment]. In K. Schweizer (Ed.), *Leistung und Leistungsdiagnostik* (pp. 127–143). Heidelberg, Germany: Springer.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London, UK: Routledge.

Haworth, C. M. A., Kovas, Y., Dale, P. S., & Plomin, R. (2008). Science in elementary school: Generalist genes and school environments. *Intelligence*, *36*, 694–701.

Hopmann, S., Brinek, G., & Retzl, M. (Eds.). (2007). *PISA zufolge PISA* [PISA according to PISA]. Wien, Austria: LIT-Verlag.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*, 3–7.

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur Test (Form 4, BIS-T4)* [Berlin Intelligence Structure Test]. Göttingen, Germany: Hogrefe.

Jahnke, T., & Meyerhöfer, W. (Eds.). (2007). *PISA & Co—Critique of a program.* Hildesheim, Germany: Franzbecker.

Jensen, A. R. (1989). The relationship between learning and intelligence. *Learning and Individual Differences*, *1*, 37–62.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Kobarg, M., & Dalehefte, I. M. (2012). Wie viel Intelligenz ist in Mathematiktests? [How much intelligence is in tests of mathematics?]. *IPN-Blätter*, *29*(2), 7.

Lohman, D. F., & Hagen, E. P. (2002). *Cognitive Abilities Test (CogAT; Form 6): Research handbook.* Itasca, IL: Riverside Publishing.

Maas, H. L. J. v. d., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842–861.

Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning.* Cambridge, UK: Cambridge University Press.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., . . . O'Connor, K. M. (2003). *TIMSS assessment frameworks and specifications 2003* (2nd ed.). Chestnut Hill, MA: Boston College.

Neubrand, M., Klieme, E., Lüdtke, O., & Neubrand, J. (2002). Degrees of competence and models of difficulty for the PISA-test of mathematical literacy. *Unterrichtswissenschaft*, *30*, 100–119.

Organisation for Economic Cooperation and Development. (2003). *PISA 2003 assessment framework. Knowledge and skills.* Paris, France: Author.

Organisation for Economic Cooperation and Development. (2009). *Take the test: Sample questions from OECD's PISA assessments*. Paris, France: Author.

Organisation for Economic Cooperation and Development. (2014). *PISA 2012 results: Creative problem solving* (Vol. V). Paris, France: Author.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? [What do international student assessment studies measure?] *Psychologische Rundschau*, *57*, 69–86.

Rindermann, H. (2007). The big *G*-factor of national cognitive ability. *European Journal of Personality*, *21*, 767–787.

Rindermann, H., Flores-Mendoza, C., & Mansur-Alves, M. (2010). Reciprocal effects between fluid and crystallized intelligence and their dependence on parents' socioeconomic status and education. *Learning and Individual Differences*, *20*, 544–548.

Rindermann, H., & Heller, K. A. (2005). The benefit of gifted classes and talent schools for developing students' competences and enhancing academic self-concept. *Zeitschrift für Pädagogische Psychologie*, *19*, 133–136.

Rindermann, H., Michou, C. D., & Thompson, J. (2011). Children's writing ability: Effects of parent's education, mental speed and intelligence. *Learning and Individual Differences*, *21*, 562–568.

Rindermann, H., & Neubauer, A. (2000). Speed of information processing and success at school: Do basal measures of intelligence have predictive validity? *Diagnostica*, *46*, 8–17.

Rindermann, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education*, *42*, 377–399.

Rost, D. H., & Sparfeldt, J. R. (2007). Reading comprehension without reading? On the construct validity of multiple-choice reading comprehension test items. *Zeitschrift für Pädagogische Psychologie*, *21*, 305–314.

Shrout, P. E. (1993). Analyzing consensus in personality judgments: A variance components approach. *Journal of Personality*, *61*, 769–788.

Steinmayr, R., & Meißner, A. (2013). The importance of intelligence and ability self-concept for the prediction of standardized achievement tests and grades in Mathematics. *Zeitschrift für Pädagogische Psychologie*, *27*, 273–282.

Thompson, J. (2013, October 11). *How illiterate is the OECD? Psychological comments*. Retrieved from http://drjamesthompson.blogspot.de/2013/10/how-illiterate-is-oecd.html

Ullrich, M., Schnotz, W., Horz, H., McElvany, N., Schroeder, S., & Baumert, J. (2012). Kognitionspsychologische Aspekte eines Kompetenzmodells zur Bild-Text-Integration [Cognitive psychological aspects of a competence model of picture-text-integration]. *Psychologische Rundschau*, *63*, 11–17.

Wang, J. (1998). A content examination of the TIMSS items. *Phi Delta Kappan*, *80*, 36–38.

APPENDIX A

Definitions of the Rating Scales

Reading Competence (Literacy). The competence to understand, use, and reflect on different forms of written text material (literature, everyday texts, tables, and figures).

Verbal Competence. Knowledge of a listener/speaker and reader/writer of his language (words/terms, grammar, and pragmatics), that is, his or her knowledge of rules enabling him or her to understand and produce correct sentences regarding content, grammar and pragmatics.

Math Competence (Literacy). The ability to reason mathematically correctly and to solve mathematical problems correctly based on math knowledge.

Science Competence (Literacy). The ability to solve scientific problems correctly and to reason in the science domain correctly based on knowledge.

Problem-Solving Ability. The ability of a person to deal with and to solve real, interdisciplinary problems whose solution is neither obvious nor based on only one area of expertise.

Reasoning. The ability to (inductively) draw the right general conclusions from several individual observations, and the ability to (deductively) apply general rules to single problems.

General Knowledge. General knowledge is the part of knowledge that everyone has or should have in order to be able to participate in social life. General knowledge refers to justifiable and relevant content.

General Intelligence. The ability of abstract thinking, of understanding, and insight in order to recognize, comprehend, and establish structures, relationships, contexts, and meanings. Intelligence is the ability to think.

Reference to Curriculum. If a task is solvable by using knowledge that is or should be acquired in the course of school instruction, then this is a curriculum-related task. In contrast, if a task is solvable without specific school knowledge (e.g., by using knowledge acquired outside school or by using information already contained in the task), then this is a curriculum-unrelated task.

Knowledge Retrieval. Preexisting knowledge has to be recalled from memory. The task can be solved by retrieving knowledge; no extensive thinking processes are required.

Thinking/Comprehension. In order to solve tasks, extensive thinking processes are required. These thinking processes are also based on preexisting knowledge that needs to be retrieved from memory or on information that needs to be extracted from the task. Cognitive operations are carried out while applying this knowledge or information.

APPENDIX B

Rating Scales From the Answering Sheets

From your point of view, how difficult is the task for 15 year olds?											
	very easy	0	1	2	3	4	5	6	7	8	very difficult
From your point of view, to what extent does the task assess the following competences?											
Reading competence	not at all	0	1	2	3	4	5	6	7	8	completely
Verbal competence	not at all	0	1	2	3	4	5	6	7	8	completely
Math competence	not at all	0	1	2	3	4	5	6	7	8	completely
Science competence	not at all	0	1	2	3	4	5	6	7	8	completely
Problem-solving ability	not at all	0	1	2	3	4	5	6	7	8	completely
Reasoning	not at all	0	1	2	3	4	5	6	7	8	completely
General knowledge	not at all	0	1	2	3	4	5	6	7	8	completely
General intelligence	not at all	0	1	2	3	4	5	6	7	8	completely
Do you think that the task is more curriculum-related or rather curriculum-unrelated?	curriculum-unrelated	0	1	2	3	4	5	6	7	8	curriculum-related
Can the correct solution be obtained rather by knowledge recall or rather by thinking/understanding?	knowledge retrieval	−4	−3	−2	−1	0	1	2	3	4	thinking/ comprehension