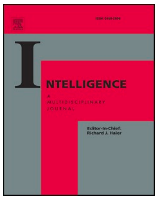
Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell



The relation between science achievement and general cognitive abilities in large-scale assessments

Nele Kampa [a,b,*], Ronny Scherer [c], Steffani Saß [d], Stefan Schipolowski [e]

[a] Leibniz-Institute for Science and Mathematics Education, Kiel, Germany
[b] University College of Education Tyrol, Institute for Research and Development in Didactics and Educational Science, Innsbruck, Austria
[c] Centre for Educational Measurement at the University of Oslo, Faculty of Educational Sciences, Norway
[d] Steffani Saß, Institute of Psychological Learning Research, University of Kiel, Germany
[e] Stefan Schipolowski, Institute for Educational Quality Improvement, Berlin, Germany

ARTICLE INFO

Keywords:
General cognitive abilities
School grades
Science achievement
Structural equation modeling

ABSTRACT

Although large-scale assessments (LSA) of school achievement claim to measure domain-specific achievement, they have been criticized for primarily measuring domain-general abilities. Numerous studies provide evidence that LSA of mathematical achievement as well as verbal achievement cover both general cognitive abilities (GCA) and domain-specific achievement dimensions. We extend previous research by analyzing a standards-oriented and literacy-oriented LSA in the domain of science to determine the relation of these two assessment types with domain-general abilities. While literacy-oriented assessments focus on the knowledge and skills students need to meet the demands of modern societies, standards-oriented assessments focus on national educational standards and curricula. A sample of 1722 students worked on three assessments: (a) the PISA scientific literacy assessment; (b) a standards-oriented assessment based on the German National Educational Standards in biology, chemistry, and physics developed by the Institute for Educational Quality Improvement (IQB); and (c) a GCA test. Comparisons of competing structural models showed that models differentiating between domain-specific achievement and GCA best represented the structure of the assessments. Furthermore, standards-oriented and literacy-oriented LSAs in science shared common variance with GCA but also comprised specific variance. In addition to a factor representing students' GCA, we identified a science literacy-oriented and two standards-oriented factors. Relations with school grades in various STEM and non-STEM subjects were mixed and only partly provided evidence for the specificity of science LSAs. Our findings are important for understanding and interpreting results of LSAs in the contexts of GCA and science. We discuss our outcomes with respect to educational monitoring practices.

1. Theoretical background

Large-scale assessments (LSA) of school achievement, such as the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), have an educational monitoring function and are therefore of high political impact (Lietz & Tobin, 2016). Besides their impact on educational policy and the development of curricula, the conceptualization of LSA also affects research on education and educational practices (Kind, 2013; Klieme, 2013). The reporting of LSA is based on the assumption that achievement scores primarily reflect domain-specific achievement—an assumption that has been challenged. Specifically, some studies have criticized the conceptualization of national and international LSA as mainly measuring broad, *domain-general* intelligence factors instead of *domain-specific* achievement, thus questioning their primary functions in the monitoring system (e.g., Koenig, Frey, & Detterman, 2008; Rindermann, 2007). If domain-specific achievement is not appropriately reflected in LSA, proficiency level models will not mirror domain-specific achievement and meaningful differences between domains such as science, mathematics, and verbal achievement. These aspects, however, represent main features of monitoring in education that need to be covered by LSA (Leutner, Hartig, & Jude, 2008). On the other hand, there are a number of studies which support the notion that LSA measure both, general cognitive abilities (GCA) and domain-specific

[*] Corresponding author at: University College of Education Tyrol, Institute for Research and Development in Didactics and Educational Science, Innsbruck, Austria.
E-mail addresses: nele.kampa@ph-tirol.ac.at (N. Kampa), ronny.scherer@cemo.uio.no (R. Scherer), sass@ipl.uni-kiel.de (S. Saß), stefan.schipolowski@iqb.hu-berlin.de (S. Schipolowski).

https://doi.org/10.1016/j.intell.2021.101529
Received 10 June 2020; Received in revised form 22 February 2021; Accepted 23 February 2021

achievement. These studies showed that domain-specific achievement constitutes a substantive part of achievement scores for mathematics and verbal achievement (e.g., Brunner, 2008; Saß, Kampa, & Köller, 2017). They applied confirmatory factor analyses (CFA) and structural equation models (SEM) to LSA data to determine the internal structure of the tested achievement domains and their relation to external criteria. If the assumption of measuring domain-specific achievement does not hold true in these analyses, the results of LSA will have to be interpreted with caution regarding their educational monitoring function.

Since much prior research focused on mathematical and verbal achievement (e.g., Brunner, 2008; Saß et al., 2017), we add new evidence to this ongoing debate by investigating the relation between GCA and achievement as measured in a standards-oriented and a literacy-oriented LSA in the domain of science. It should be noted that the curricula of all federal states have been adapted to reflect the educational standards. Hence, standards-oriented LSA can be considered largely congruent to curriculum-oriented LSA, such as TIMSS, as we will discuss in the following section. With GCA, we refer to general cognitive abilities in terms of the *g* factor established in many contemporary theories of intelligence structure (e.g., Carroll, 1993, 2003; McGrew, 2009), which is an overarching general factor at the top of the ability hierarchy involved in all kinds of cognitive performances. The *g* factor has been characterized by Carroll (1994, p. 62) as having "its highest loading for factors and variables that involve the level of complexity at which individuals are able to handle basic processes of induction, deduction, and comprehension". In line with this definition, we also consider reasoning ability (gf) a domain-general ability factor. Not only has it been argued that gf and *g* are closely related, some authors even consider both factors to be equivalent (e.g., Gustafsson, 1984; Undheim & Gustafsson, 1987).

Our study enriches the debate on what LSA measure in several ways: First, focusing on science achievement answers the call of researchers in science education to direct the focus on science assessment (Songer & Ruiz-Primo, 2012). Second, to our best knowledge, this is the first study to administer both assessment types to the same sample of students. With this approach, we eliminate possible biases, which could occur when simply comparing results from different studies on the basis of samples that show demographic or geographic differences (e.g., gender, poverty rate). Moreover, this feature of our study attempts to fulfill the demand for coherence within the science education assessment system (Pellegrino, 2012). We hope that our findings will prove crucial to valid reporting and interpretation in the context of educational monitoring in science education. By incorporating both assessment types, we are able to identify differences and similarities of these assessment types concerning their relations to GCA, as measured by cognitive ability tests. Third, we investigate national and international science assessments in Germany to explore how the constructs that underlie these LSA could be interpreted and, ultimately, reported. In that way our findings contribute to the crafting of a validity argument (Pellegrino, DiBello, & Goldman, 2016).

1.1. Standards-oriented and literacy-oriented science assessment

In LSA, science achievement, like achievement in other domains, is regularly monitored through either literacy-oriented or curriculum-oriented assessments (Wagemaker, 2014). Both cover a broad range of students' knowledge and understanding of science. A synopsis on similarities and differences between these two approaches for mathematics has previously been discussed (e.g., Klieme, 2016; Wu, 2010). Klieme (2016) concluded that even though both approaches are "indicators of overall achievement in mathematics" (p. 9), curriculum-oriented LSA reflect the implemented curricula to a greater extent. A rating study by Rindermann and Baumeister (2015) on similarities between PISA (literacy-oriented) and TIMSS (curriculum-oriented) LSA across all domains revealed that the TIMSS tasks are rated as more curriculum-related than the PISA tasks. We therefore take a closer look at the conceptualizations of curriculum−/standards-oriented approaches in comparison with literacy-oriented approaches.

TIMSS or the National Education Standards assessments provide examples of mainly curriculum-oriented science LSA. In TIMSS, the intended, implemented, and attained curricula form the basis of the test items (Mullis & Martin, 2013). These curricular aspects reflect a conceptualization of science "that students are expected to learn as defined in countries' curriculum policies and publications […]" (Mullis & Martin, 2013, p. 4). The assessment covers general topics such as human health, chemical change, light and sound, and earth structure; furthermore, it includes physical features within the content domains biology, chemistry, physics, and earth science and within the cognitive domains knowing, applying, and reasoning. This differentiation between and within mental representations and processes complies with the criteria of a profound model of learning for any assessment (Pellegrino, 2012).

The normative framework of the National Educational Standards aims to monitor the extent to which academic performance at certain stages meets the proficiency expectations formulated in the educational standards (Stanat, Schipolowski, Mahler, Weirich, & Henschel, 2019). For example, the German National Educational Standards in biology, chemistry, and physics define proficiency levels students should have reached by the end of lower secondary level education (tenth grade, equivalent to a junior high school graduation; Stanat et al., 2019). The standards contain content areas specific to science knowledge and skills: content knowledge, scientific inquiry, argumentation, and evaluation (Stanat et al., 2019). These content areas are further broken down. For example, content knowledge in biology is differentiated into the three basic concepts: systems, structure and function, and development. Scientific inquiry is differentiated into three main processes: scientific investigations, scientific modeling, and scientific theorizing. The standards-oriented LSA does not aim to examine whether the assessment content is part of the curriculum but rather to investigate whether schooling leads to meeting normative standards. This approach harmonized the curricula of the federal states and the National Educational Standards in Germany to some extent (e.g., Hessisches Kultusministerium (Hrsg.), n.d.).

When compared to other countries, for example the United States of America, the conceptualization of the German National Standards shows similarities as well as differences (National Research Council [NRC], 2014). In the USA, the three major dimensions of the framework are scientific and engineering practices, cross-cutting concepts, and disciplinary core ideas. The core ideas (equivalent to the basic concepts in the German Educational Standards) are differentiated further within and the scientific practices across the science disciplines. In the German Educational Standards, the cross-cutting concepts are not defined.

Literacy-oriented LSA, such as PISA, focus on the knowledge and skills students need to meet the demands of modern societies outside the classroom (Kind, 2013). Scientific literacy, as defined in the PISA framework, refers to:

> […] an individual's scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena and to draw evidence-based conclusions about science-related issues, understanding of the characteristic features of science as a form of human knowledge and enquiry, awareness of how science and technology shape our material, intellectual, and cultural environments, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen.
>
> (Organisation for Economic Co-operation and Development [OECD], 2013, p. 17)

According to this definition, scientific literacy is differentiated into *knowledge of science* and *knowledge about science*. Since PISA 2015, the latter has been further differentiated into procedural and epistemic knowledge (OECD, 2016). Knowledge of science includes major fields within the disciplines of physics, chemistry, biology, earth, and space science as well as science-based technology. Biology (living systems) like the other major fields is further differentiated into cells, humans, populations, ecosystems, and biosphere. *Knowledge about science* includes scientific enquiry and scientific explanations. Scientific enquiry contains origin, purpose, experiments, data type, measurement, and characteristics of results, whereas scientific explanations contains types, formation, rules, and outcomes.

As our description shows, the main differences between the assessment approaches lie in the content they cover and the purpose they serve within the monitoring system. Some researchers argue that the cognitive processes underlying both assessment types are similar (e.g., Rindermann & Baumeister, 2015; Saß et al., 2017). In fact, all science LSA frameworks specify a knowledge and an inquiry dimension (Kind, 2013). This similarity across frameworks implies high conceptual and empirical congruency among standards-oriented and literacy-oriented assessment approaches. At the same time, specific differences between the knowledge and inquiry dimensions may exist in their conceptualization and implementation in the corresponding test items. Moreover, it was within the debate on domain-specificity and domain-generality of LSA reporting, that the question of differences between these assessment approaches arose (Klieme, 2016). From a conceptual point of view, the constructs LSA intend to measure and the constructs representing GCA share both similarities and differences—this observation may manifest in empirical evidence backing their substantial correlation yet not their equivalence.

1.2. Science achievement and general cognitive abilities

To our best knowledge, only one study investigated the interrelations between science achievement as measured in LSA and GCA (Kampa & Köller, 2016). In the study by Kampa and Köller (2016), GCA is represented by a domain-general gf factor based on a figural reasoning test. The study found that science achievement is empirically distinct from domain-general abilities —in this case, gf—and that both constructs are significantly correlated (r = .65). However, this relation was only investigated for a standards-oriented LSA. A study on argumentation in science revealed relations between $r_{\text{numerical reasoning}} = .50$ and $r_{\text{figural reasoning}} = .56$ with cognitive abilities (Heitmann, Hecht, Schwanewedel, & Schipolowski, 2017). A rating study focusing on the domains reading, mathematics, science, and problem solving showed that the curriculum-oriented TIMSS tasks were judged to require less intelligence and more curriculum-related knowledge than the literacy-oriented PISA tasks (Rindermann & Baumeister, 2015). We also consulted the results of these studies on the relations between mathematical achievement in LSA and GCA (Brunner, 2008; Saß et al., 2017). Specifically, the studies on GCA and mathematical as well as verbal achievement investigated the internal structure of LSA by specifying models that were based on factor-analytic research on the structure of GCA (Baumert, Brunner, Lüdtke, & Trautwein, 2007; Brunner, 2008; Carroll, 1993; Cattell, 1963; Horn & Noll, 1997; McGrew, 2009; Saß et al., 2017; Spearman, 1904)—these models are shown in Fig. 1.

The first model conceptualizes a global *g*, a general mental ability necessary for successful performance across all domains (see Fig. 1a). Other models typically define achievement as measured by LSA as correlated-factors and hierarchical models. The correlated-factors model (Gignac & Kretzschmar, 2017) postulates several interrelated factors—in the educational context, domain-general factors such as GCA and domain-specific factors such as science achievement, math achievement or verbal ability—but no overarching *g*-factor (see Fig. 1b). In these types of models, which are based on gc-gf-theory (Cattell, 1963; Horn & Noll, 1997), gf refers to individual differences in the ability to reason, which is closely related to *g* (Carroll, 1993; Undheim & Gustafsson, 1987). The third model postulates a hierarchical structure with several specific factors on a first level, more general factors on a second level and a general, domain-independent factor *g* on a third level. A specific implementation of these hierarchical models is the nested-factor model (see Fig. 1c; Gustafsson & Balke, 1993). In this model, the common variance of all items is explained by a *g*-factor and additional variance is explained by one or several specific factors. These factors can be domain-specific or/and domain-general factors. Since these specific factors are already controlled for the variance explained by *g*, they cannot be interpreted in the same way as the domain-specific factors in correlated-factors or other hierarchical models. We build on these three

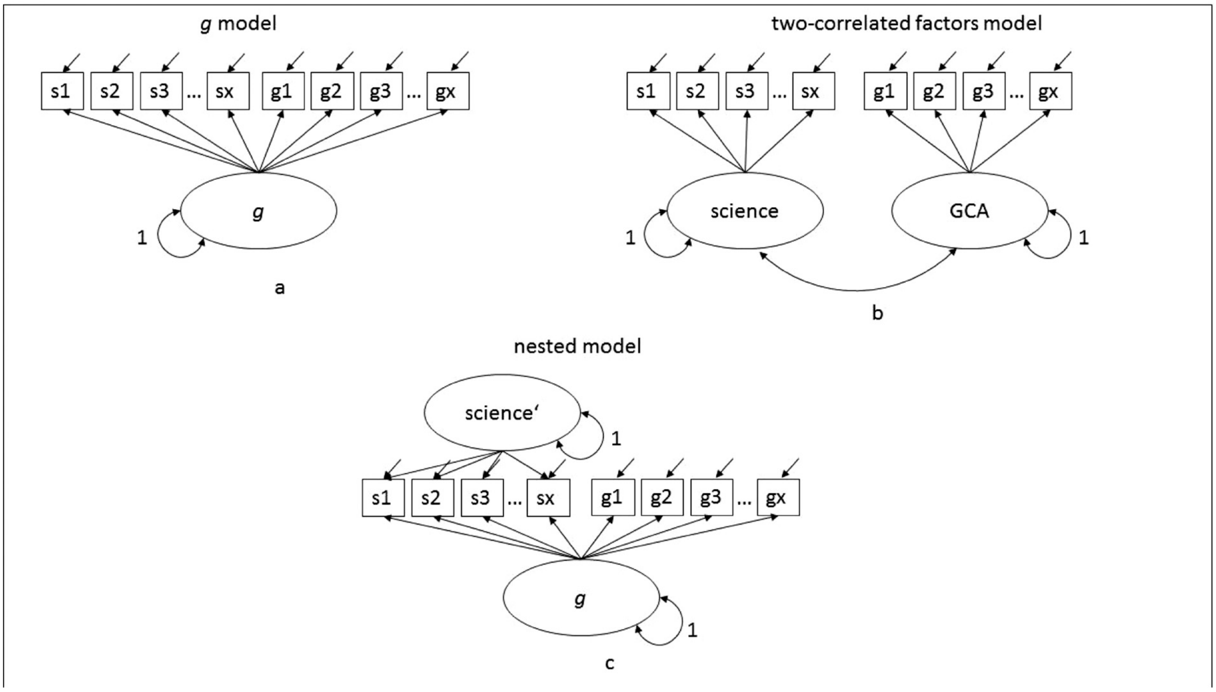


Fig. 1. a to c Models of the relation between domain-specific achievement and GCA. s = manifest indicators for domain-specific achievement, g = manifest indicators for GCA, *g* = general cognitive ability, Gf = fluid general cognitive ability.

conceptualizations—*g*-factor, correlated factors, and hierarchical factors—to investigate the relations between science achievement and domain-general GCA.

In previous studies on GCA and domain-specific achievement, a correlated-factors and/or a nested-factor model showed good model fit and was superior to the single-factor *g* model (Baumert et al., 2007; Brunner, 2008 ; Saß et al., 2017). Several authors found correlations between r = .38 and .89 for mathematics achievement and GCA (Brunner, 2008; Saß et al., 2017) and of $r = .83$ for verbal achievement and GCA (Brunner, 2008). In the context of our study, the investigation by Saß et al. (2017) needs to be pointed out in particular. The authors specified and estimated three models to different curricular- and literacy-oriented LSA for mathematical achievement of different cross-sectional studies. They applied a *g*-factor model, a correlated-factors model that assumes several correlated domain-specific and domain-general abilities but no overarching factor, and a nested-factor model assuming that abilities can be differentiated into GCA and domain-specific achievement. Across all grade levels and across the two assessment types, the correlated-factors model and/or the nested-factor model, both of which represent a conceptualization containing several cognitive factors, showed the best model fit. For mathematical achievement, these results led to the conclusion that LSA measure more than just general intelligence represented by *g*. The authors argued that the major shortcoming of their study was the cross-sectional nature of the investigation, which also meant using different samples for curricular-oriented and literacy-oriented LSA, respectively. Since we administered the two assessment types to one sample, this shortcoming can be eliminated in the present study.

1.3. Relations among school grades, science achievement, and GCA

When determining the distinctness of domain-specific achievement as compared to domain-general GCA, researchers have employed external criteria. These criteria describe the nature of the factors representing certain constructs. Within intelligence research, relating cognitive abilities to school performance constitutes achievement evidence (Horn & Noll, 1997). In previous studies on domain-general and domain-specific factors, external criteria such as school grades, gender or domain-specific motivational variables, such as self-concept, have been considered (Brunner, 2008; Saß et al., 2017; Schipolowski, Wilhelm, & Schroeders, 2014). Additionally, studies on the prediction of domain-general GCA and domain-specific achievement by these criteria serve as an indicator of their differential relations with external criteria (e.g., Wee, 2018; Ziegler & Peikert, 2018).

Our review of the extant literature resulted in only one study that focused on the relation between science grades and science achievement. In a study by Kampa (2012), the relation between a standards-oriented LSA in biology and grades in biology was $r = .26$ for content knowledge and $r = .25$ for scientific inquiry. As studies on this relation within the science domain(s) are rare, we again consult results from mathematics and verbal achievement. The correlation between mathematics assessment and the school grade in mathematics for both assessment types ranged between $r = .24$ and $r = .49$. In addition, school grades in German (first language) were less related to domain-general GCA than to verbal achievement. Correlations in nested-factor models, i.e. correlations between school grades and a domain-specific factor controlled for intelligence, as represented by the *g* factor, were lower than in correlated-factors models and ranged between r = .12 and .39. The corresponding correlation with GCA ranged between $r = .22$ and .33 (Brunner, 2008; Saß et al., 2017).

The small differences between science, mathematics, and verbal achievement correlations might indicate that domain-general GCA is more closely related to mathematics achievement than to science and verbal achievement. Much prior research focuses on the investigation of relations between achievement and grades in one domain; differential relations between grades in multiple subjects and achievement in one specific domain are rarely scrutinized. These few studies found that within-domain correlations are higher than correlations across domains (Brunner, 2008). A previous study on LSA in biology revealed the same relations for content knowledge and scientific inquiry to school grades in German, mathematics as well as the three science domains of biology, chemistry, and physics (Kampa, 2012).

In the present study, we transfer previous results from studies in the domains of mathematics and verbal achievement to literacy- and standards-oriented assessment in the domain of science. So far, a comparison between these two assessment approaches has not been carried out.

1.4. Aims of the study and hypotheses

In light of the monitoring function of LSA and its impact on research in education, we aim to provide evidence for the domain-specificity of literacy-oriented and standards-oriented science LSA. By providing new evidence on the relation between domain-general and domain-specific cognitive abilities, our study also contributes to intelligence research and takes up the debate about the role of *g* in education (e.g., Baumert et al., 2007; Brunner, 2008; McGrew & Wendling, 2010). In the following, we will formulate our hypotheses, which also take into account earlier results from mathematics and verbal LSA as we cannot rely on studies from the field of science education alone. The first three hypotheses are based on evidence for science LSA measuring domain-specific achievement as well as domain-general abilities. Note that since our analyses will be conducted using Item Response Theory (IRT, see Statistical Analyses), we apply the IRT nomenclature (i.e., we speak of “dimensions” instead of “factors”).

H1. Models containing several domain-specific dimensions as well as a domain-general dimension (correlated-dimensions and nested model) show better fit to the data than a one-dimensional *g* model (congruent to g-factor model).

H2. The correlations between science achievement (as measured in standards-oriented and literacy-oriented science LSA, respectively) and GCA (other domain correlation) are lower than the correlations between the measures of science achievement (i.e. the two assessment types standards-oriented and the literacy-oriented; same domain correlation).

H3. The correlation between standards-oriented and literacy-oriented science achievement as measured in LSA remains significant after controlling for a broad domain-general intelligence that is represented by *g* (nested model).

The first hypothesis focuses on the structure of science LSA. If this hypothesis is supported by our data we can assume that science LSA reflect both domain-specific achievement and domain-general abilities. The second and third hypotheses are phrased to further support this evidence. Since the two science LSA claim to measure very similar constructs with different assessment approaches, they should be more closely related to each other than to GCA. Moreover, if both assessment approaches measure science achievement over and above GCA, they should still be related once the variance stemming from domain-general intelligence is accounted for.

Previous studies on mathematics achievement showed that same-domain grades are more closely related to achievement than other-domain grades. We therefore expect the same patterns regarding the correlation of science grades in contrast to grades in German and mathematics with science LSA. Our two hypotheses on the relations to school grades as external criteria consequently read:

H4. Science grades correlate more strongly with science achievement than with GCA. This differential correlational pattern will not show for the other-domain grades (i.e., grades in mathematics and German).

H5. The correlation between science grades and both science achievements measured in LSA will remain significant when controlling for a broad domain-general intelligence factor which is represented by *g* (nested model).

In our theoretical background, we briefly elaborated on the two subdimensions of scientific literacy, namely content knowledge and scientific inquiry. This differentiation is part of the conceptualizations of science achievement in LSA. However, as our investigation does not focus on this distinction, we did not formulate hypotheses concerning these subdimensions.

2. Methods

2.1. Sample and procedure

We used data of a study that took place in 80 schools in six federal states of Germany in May 2012 and was carried out by the Institute for Educational Quality Improvement (IQB; Pant et al., 2015). The schools participated voluntarily and the sample consisted of 1730 students in 9th grade (49.3% female). The students were enrolled in all German school types (40% academic secondary schools [*Gymnasium*], 60% non-academic schools). They completed a PISA science test, a German National Educational Standards in Science assessment (Pant et al., 2013), further assessments not considered in our analyses and a short questionnaire. The students worked on these papers on two consecutive days within a three-hour period on each day and were supervised by trained test administrators. We used the subsample of 1722 students who worked on the science PISA-test and a test on the National Educational Standards for Science (Lenski et al., 2016).

2.2. Measures

Science education researchers have argued that in order for science assessment to be of high quality, science educators, science education

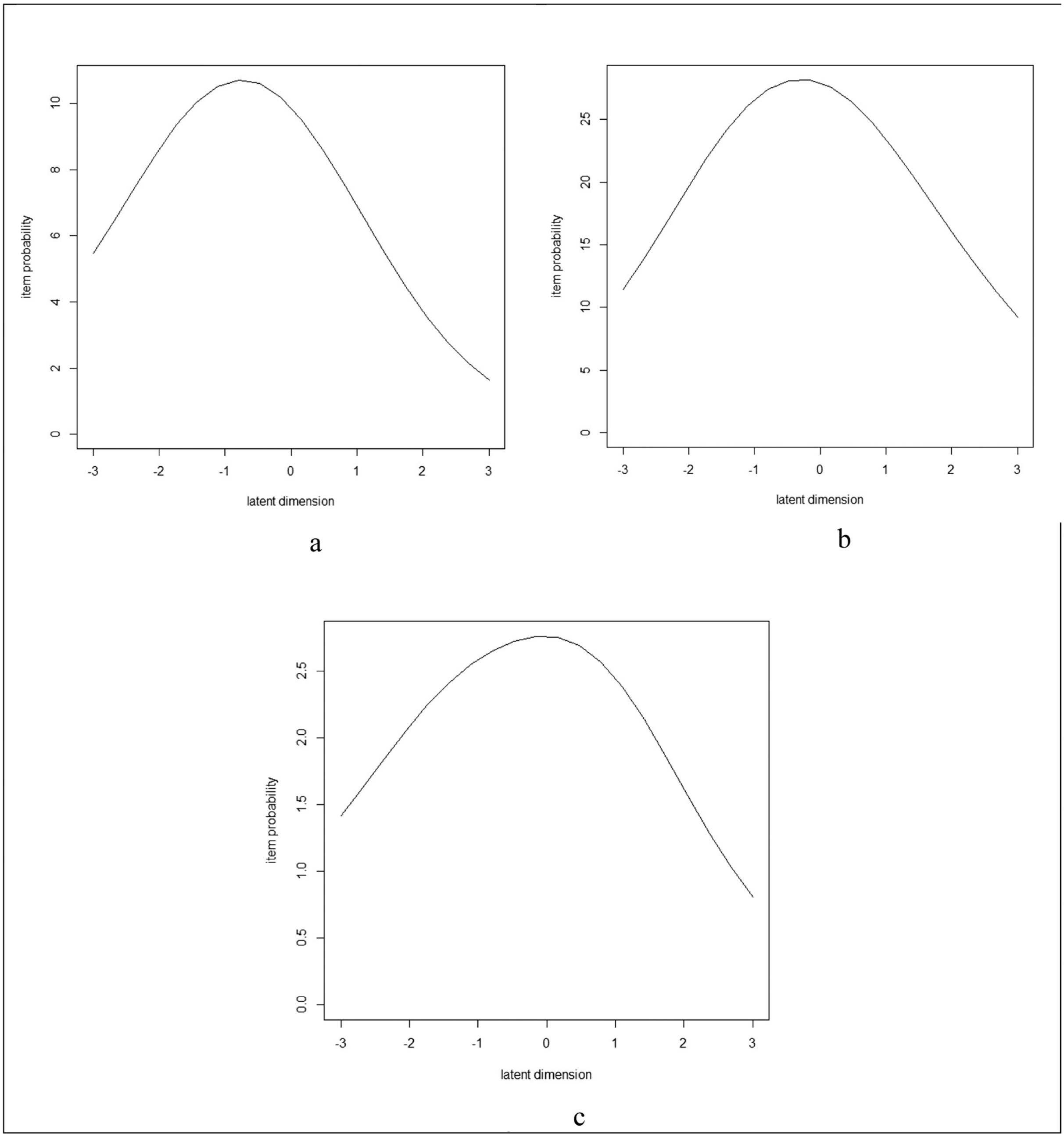


Fig. 2. a to c Test information curves for the (a) literacy-oriented and (b) standards-oriented LSA as well as (c) the general cognitive abilities test.

researchers, psychometricians, and language specialists need to be part of the development process (Songer & Ruiz-Primo, 2012). The international, literacy-oriented LSA was monitored and supported by an international expert group and developed through contracts with external partners (OECD, 2014a). The national, standards-oriented LSA test was developed through the collaboration of researchers and practitioners from diverse academic backgrounds (Stanat et al., 2019).

2.2.1. Literacy-oriented science LSA

The PISA test was developed in a multi-stage process by nine renowned research institutes around the world and translated into German (OECD, 2014a). We used the 53 items administered in Germany in 2012. The items were tested in a field study, during which items that did not show appropriate fit were eliminated. The items were administered in four different response formats: open response (17 items), multiple choice (17), complex multiple choice (18), and closed constructed response (1). They fall into two categories, the competencies component and the knowledge component. Regarding the competencies component, using scientific evidence is represented by 18 items, identifying scientific issues by 13, and explaining phenomena scientifically by 22 items. Regarding knowledge component, the aspect of knowledge of science is represented by 26 items (physical systems 6 items, earth and space systems 7, living systems 9, technical systems 4), the aspect of knowledge about science is represented by 27 items (scientific explanations 13 and scientific enquiry 14 items). The test information curve is displayed in Fig. 2a. Since not all students could answer all items within the given test time, the PISA items were assigned to 3 testlets of 20 min. Each student worked on one testlet.

2.2.2. Standards-oriented science LSA

The National Educational Standards assessment was developed under the supervision of the IQB. The test development comprised several stages (Stanat et al., 2019). First, science teachers developed items based on an item development model, which was constructed by science education researchers in biology, chemistry, and physics. The items were evaluated by (different) science education researchers and a language specialist and tested for comprehension problems in the classes of the science teachers. Second, the items were pre-piloted, piloted in five federal states of Germany, and then normed on a representative sample of 13,328 students. During these three stages, items that showed inappropriate fit values were dropped. This development process resulted in an item pool for the monitoring of the National Science Standards in content knowledge and scientific inquiry, which takes place every six years.

In the present study, we administered 147 items of this National Educational Standards assessment, on biology (58 items), chemistry (50), and physics (39). Seventy-five items can be allocated to content knowledge and 72 items to scientific inquiry. The test information curve is displayed in Fig. 2b. The standards-based items were assigned to 12 testlets. Each student worked on 3 testlets of 20 min.

2.2.3. GCA and school grades

Afterwards, the students worked on the figural reasoning scale of the Berlin Test of Fluid and Crystallized Intelligence (BEFKI; Schipolowski et al., 2014). The scale is comprised of 16 items; each item consisted of a sequence of geometric shapes whose elements changed according to implicit rules. To complete the tasks, students had to infer these rules and choose the next two shapes in the sequence from a number of given alternatives. The scores resulting from this assessment of the figural aspect of fluid intelligence can be regarded as a proxy for individual differences in fluid intelligence, which is strongly related to *g* (Carroll, 1994; Wilhelm, 2004). Note that figural reasoning tests are widely considered as marker tests for *g* (Jensen, 1998). The test information curve is displayed in Fig. 2c. Finally, the school officials reported students' grades in German, mathematics, biology, chemistry, and physics on their latest report card (Lenski et al., 2016). In Germany, grades range from 1 to 6 with 1 representing the highest grade (*superior*).

2.3. Statistical analyses

All analyses were calculated based on maximum likelihood estimation with robust standard errors (MLR) in the statistical software package Mplus 7.4 (Muthén & Muthén, 1998-2012). We proceeded in three steps. In a first step, we applied measurement models for each science LSA separately. In particular, we calculated single- and correlated-dimensions models representing the different assumptions on the internal structure of the measured constructs. In all models, the manifest indicators corresponded to the individual test items. Furthermore, we constrained the discrimination parameters on each dimension to be equal in all models. As a result of choosing this type of constraint, our applied models can be attributed to the framework of IRT and they are comparable to CFA with categorical indicators (Brown, 2015). Our approach had two objectives: First, we aimed at a Rasch model, a specific IRT model, to which the items were constructed to fit (for the Educational Standards see Stanat et al., 2019; for PISA see OECD, 2014b). In order to treat the three assessments equally and because it has been applied before to the reasoning scale, we also used this procedure to model GCA (Schroeders, Schipolowski, & Wilhelm, 2015). Second, we attempted to reduce the complexity of the models. Since we used the individual test items as indicators for the latent dimensions in all models, the number of parameters to be estimated was quite large, which made the models highly complex. Constraining the discrimination parameters for each measure to equality reduced the number of estimated parameters. It should be noted, however, that the IRT modeling approach provides relative but not absolute fit indices.

For the comparisons between the measurement models, we relied on the comparative fit indices Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and corrected Akaike Information Criterion (cAIC). In order to test the superiority of the correlated-dimensions model and the nested model in comparison to the *g* model, we performed the Satorra-Bentler scaled chi-square difference test (Satorra & Bentler, 2010). The lower the values of these fit indices, the better the model represents the underlying data (Wilson & De Boeck, 2004). We used the analytic option TYPE = COMPLEX to account for the nested data structure (students within classroom within schools) and treated missing values with the full information maximum likelihood (FIML) procedure.

In the one-dimensional measurement models for the literacy- and standards-oriented assessment, all science items load on one dimension for each LSA and for GCA respectively. In the two-dimensional measurement models for both LSA—the literacy-oriented and the standards-oriented—the items on content knowledge/knowledge of science load on one dimension and the items on scientific inquiry/knowledge about science on a second dimension. In the three-dimensional model for the standards-oriented LSA, the items on biology, chemistry, and physics load on one dimension each. In the six-dimensional model for the standards-oriented LSA, a content knowledge dimension and a scientific inquiry dimension is modeled for each of the three science domains (2 × 3 dimensions).

In a second step, we calculated a series of models in order to shed light on hypotheses H1 to H3 (see Fig. 3a to c). In model a – the *g* model – all items of the standards-oriented LSA, of the literacy-oriented LSA as well as those of the GCA test load on a single dimension (see Fig. 3a). In this model, *g* represents a general dimension that comprises the common variance of the science LSA and the GCA test. In the second model (b) – the correlated-dimensions model – each assessment is represented by one dimension and the dimensions are correlated with each other (see Fig. 3b). In other words, this model presumes several correlated dimensions on the same level. All dimensions in model b representing science assessments incorporate both specific and shared (*g*-related) variance to a certain degree. As we have set the discrimination parameters to be equal, the *g* model and the correlated-dimensions model are

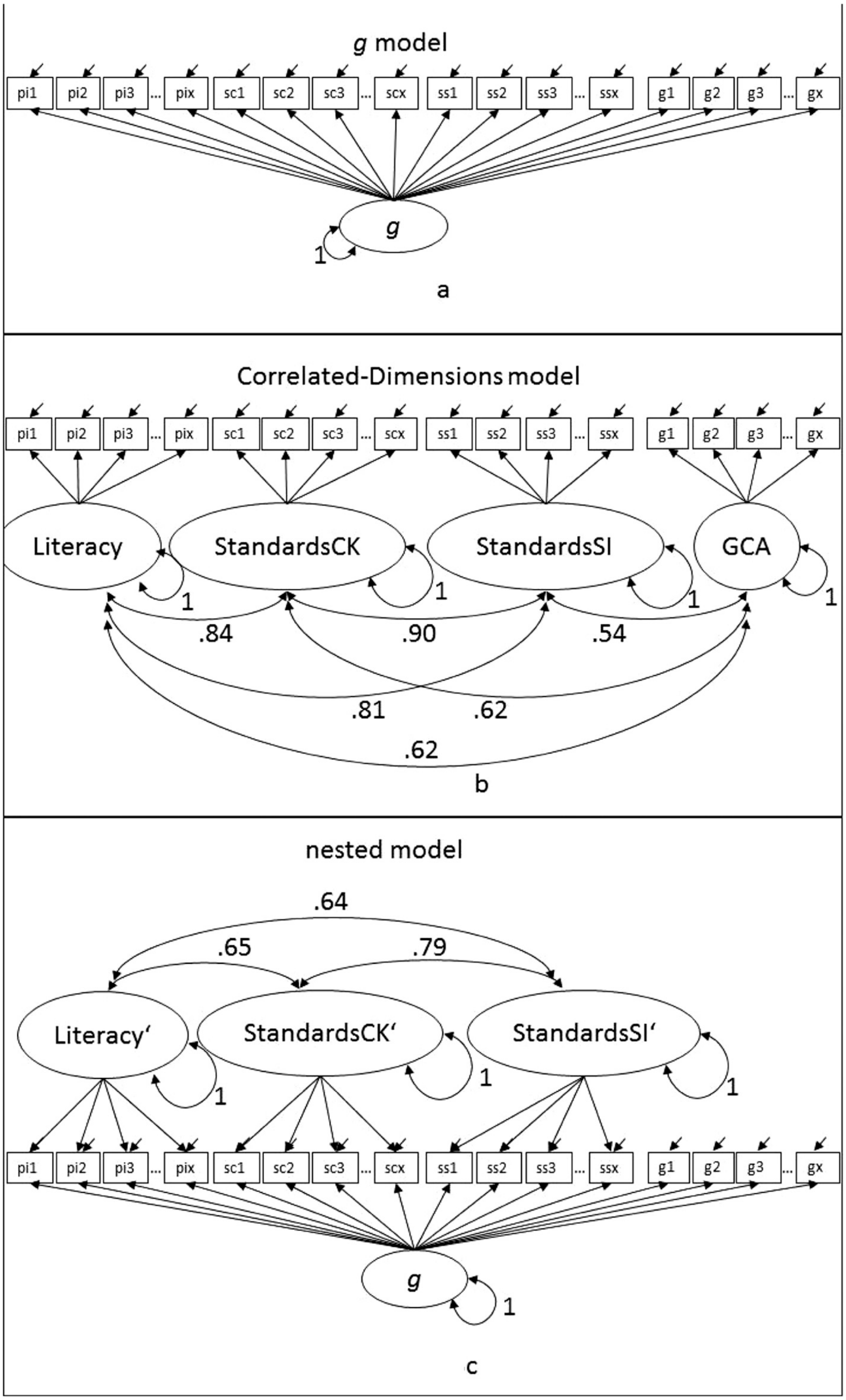


Fig. 3. a to c. The three models of domain-specific and general cognitive abilities. 3a = *g* model, 3b = correlated-dimensions model, 3c = nested-dimension model. *g* = general cognitive abilities; CK = content knowledge; SI = scientific inquiry; GCA = general cognitive abilities; sc = Standards content knowledge; ss = Standards scientific inquiry; g = manifest indicator for GCA (BEFKI).

nested within each other. In model c—the nested model—domain-specific variance is represented by a separate dimension for each construct, whereas the shared variance of all items is represented by a *g*-dimension (see Fig. 3c). The removal of shared variance of the domain-specific dimensions due to the *g*-dimension is depicted by the symbol ‘following the denomination of the dimension. Similar to model a, the *g* in model c represents a general dimension that comprises the common aspects of the science LSA and the GCA test. However, the specific dimensions are modeled to be orthogonal to *g* (i.e., the correlation between *g* and the specific dimensions is fixed to 0) and they are controlled for the common, shared (*g*-related) variance. This model stands in the tradition of hierarchical models of cognitive abilities.

We performed Wald tests (Bollen, 1989) on the correlated-dimensions model to test whether the correlations between the

Table 1
Fit indices for the measurement models for the literacy and the curricular assessment as well as for GCA.

Assessment	Model	Npar	Log-likelihood	AIC	BIC	cAIC
Literacy-oriented	One-Dimensional	55	−19,336	38,781	39,079	39,134
($n = 1769$)	Two-Dimensional	61	−19,335	38,781	39,085	39,141
Standards-oriented	One-Dimensional[2]	148	−18,931	38,160	38,869	39,018
($n = 862$)	Two Dimensional	150	−18,909	38,118	38,832	38,982
	One-Dimensional[3]	150	−18,932	38,164	38,878	39,028
	Three-Dimensional	153	−18,906	38,117	38,846	38,999
	One-Dimensional[6]	154	−18,916	38,139	38,872	39,026
	Six-Dimensional	174	−18,879	38,106	38,934	39,108
GCA	One-Dimensional	32	−13,671	27,376	27,477	27,484
($n = 1639$)						

Note. GCA = general cognitive abilities; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; cAIC = Corrected Akaike Information Criterion, Npar = Number of parameters.

[2] Discrimination parameters set equal like in the two-dimensional model.

[3] Discrimination parameters set equal like in the three-dimensional model.

[6] Discrimination parameters set equal like in the six-dimensional model.

domain-specific dimensions and the GCA dimension differ in a statistically significant way from the correlations between the domain-specific dimensions.

In a third step, we extended the models by incorporating structural relations between domain-general and domain-specific dimensions and grades in German, mathematics, biology, chemistry, and physics in the correlated-dimensions and the nested model (H4 and H5). School grades are affected by environmental factors (Südkamp & Möller, 2009; Zeidner & Schleyer, 1998). These influences could moderate the correlations between school grades on the one hand and domain-general and domain-specific dimensions on the other. In order to check the robustness of the findings on various school grades, we therefore ran all analyses separately for the subgroups academic secondary schools and non-academic secondary schools[1].

3. Results

3.1. Measurement models

For the literacy-oriented assessment, analyses resulted in a one-dimensional measurement model (see Table 1). All fit indices slightly increased in the two-dimensional solution or remained the same. For the standards-oriented assessment, the fit indices showed that the two-dimensional solution fitted the data best. While the non-sample independent BIC and the sample dependent cAIC decreased from the one-dimensional to the two-dimensional model and then increased in the following models, the sample dependent AIC continuously decreased. The correlation between the content knowledge and scientific inquiry dimensions in the two-dimensional model was $r = .89$, which was significantly different from one; Wald-$\chi^2(1) = 31.37$, $p < .001$. Therefore, we decided to represent the structure of the standards-oriented assessment with a two-dimensional measurement model.

3.2. Science achievement and general cognitive abilities (GCA)

Our first three hypotheses targeted the structure of science achievement in LSA and GCA as well as the relations among these constructs. Therefore, we first report on structural models including science achievement, GCA, and an (overarching) g-dimension (see Table 2, see Table A2 in the Appendix for a global *g* model that includes all indicators of our study).

Depending on the fit statistic, either the correlated-dimensions model or the nested model showed the best fit. The indices that account for the sample size favored the correlated-dimensions model. The standardized discrimination parameters are displayed in Table 3.

The Satorra-Bentler scaled chi-square difference test (see Table 2) shows—as we expected—that both models with domain-specific dimensions were superior to the *g* model. The results imply that science LSA, just like math and verbal LSA, measure more than just GCA.

The correlations between the dimensions within the correlated-dimensions model give a first hint on the differential relations between domain-specific achievement and domain-general abilities (see Fig. 3b). First, the correlations between the two domain-specific science dimensions on the one hand and the domain-general dimension (GCA) on the other were lower than the correlations between the domain-specific dimensions. The Wald tests for equality of the correlation pairs domain-specific dimension and domain-general dimension versus domain-specific and domain-specific dimension corroborated this picture (see Table 4).

The correlations between the literacy-oriented LSA and the standards-oriented LSA were statistically different from the correlation between the literacy-oriented dimensions and GCA. The same comparisons for the standards-oriented LSA also became significant for content knowledge and scientific inquiry. To sum up, the correlations between the domain-specific dimensions and the domain-general GCA dimension in the correlated-dimensions model reveal the same pattern of results for all measures, which is in line with our second hypothesis. Furthermore, the correlations between the domain-specific dimensions in the nested model were statistically significant as well (*ps* < .001; see Fig. 3c).

3.3. Science achievement, GCA, and school grades

The correlations with grades in various school subjects and the dimensions in the correlated-dimensions model shed some light on the specificity of the mapped abilities (see upper part of Table 5 for correlations and upper part of Table 6 for Wald-Tests; please refer to Table A1 in the Appendix for the correlations between different science dimensions and individual school grades, and to Table A3 in the Appendix for the correlations between the science dimensions and a general school grade factor). Looking at the science grades, the picture partly supports our hypothesis for biology and physics. While the relations mirror our hypothesis for the comparison of the literacy-oriented assessment versus GCA for all science grades, all relations of the physics grade and domain-specific dimensions are significantly higher than the relation of the physics grade with GCA. This finding is only true for academic secondary schools. Hence, we could only fully corroborate our hypothesis for the science domain physics in academic secondary schools.

Against our expectations, we found a corresponding differential pattern for the grade in German and again in academic secondary schools only. The relations of the grade in German with domain-specific dimensions were also significantly higher than the relation of the grade in German with GCA. As anticipated, in relation to the mathematics

[1] Due to the complexity of the models, we were not able to perform multilevel models to account for reference effects in the classroom.

Table 2
Fit indices for the models of the internal structure of the domain-specific and domain-general LSAs.

Model	Npar	Log-likelihood	AIC	BIC	cAIC	χ^2-test
g	220	−51,821	104,081	105,281	105,621	–
Correlated-Dimensions	226	−51,481	103,415	104,647	104,873	$\chi^2(6) = 1263.92$, $p < .05$
Nested	438	−51,147	103,170	105,558	105,996	$\chi^2(218) = 603.25$, $p < .05$

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; cAIC = Corrected Akaike Information Criterion, Npar = Number of parameters; *g* = general dimension; χ^2-test = Satorra-Bentler scaled χ^2-test; test calculated against the g model. Due to the large number of indicators (216) and the IRT approach, absolute fit indices (e.g., CFI, SRMR) cannot be provided.

Table 3
Standardized discrimination parameters for the models of the internal structure of the domain-specific and domain-general LSAs.

Model	Construct			
	Literacy oriented LSA	Standards-oriented LSA (science content)	Standards-oriented LSA (scientific inquiry)	GCA/*g*
g	.43	.47	.49	.34
Correlated-dimensions	.46	.48	.51	.45
Nested	.32 (0.02)	.34 (0.02)	.38 (0.02)	.35 (0.13)

Notes. LSA = large-scale assessments; GCA = general cognitive abilities; *g* = general dimension. Discrimination parameters in the IRT framework are the equivalent to factor loadings in factor analysis. In the *g*-model and correlated-dimensions model, all discrimination parameters are constraint to be equal. Therefore, we only provide one discrimination parameter per dimension. Discrimination parameters of the nested-dimension model are mean discrimination parameters and the respective standard deviations; All discrimination parameters of the nested model except 29 out of 216 loadings on the *g* dimension in the nested-dimension model were significant.

Table 4
Wald-test statistics on the tests of equality of correlation pairs in the correlated-dimensions model.

Comparison	$\chi^2(1)$	*p*-value
L-O/GCAvsL-O/S-O-CK	31.32	<.001
L-O/GCAvsL-O/S-O-SI	40.37	<.001
S-O-CK/GCAvsS-O-CK/S-O-SI	50.76	<.001
S-O-CK/GCAvsS-O-CK/L-O	28.09	<.001
S-O-SI/GCAvsS-O-SI/L-O	49.11	<.001
S-O-SI/GCAvsS-O-SI/S-O-CK	83.11	<.001

Notes. L-O = literacy-oriented, S-O- = standards-oriented, GCA = general cognitive abilities, CK = content knowledge, SI = scientific inquiry.

grade, we did not find a differential pattern regarding grades on the one hand and domain-specific as well as domain-general abilities on the other hand. An exception is the relation of the mathematics grade with the literacy-oriented dimension as compared to the relation of the same grade with the GCA dimension in non-academic secondary schools. The latter relation is statistically higher than the former.

The picture changes when looking at the same patterns for the domain-specific dimensions after removing *g*-related variance from the domain-specific science dimensions and the domain-general GCA dimension in the nested model (see lower part of Table 5 for correlations and lower part of Table 6 for Wald-tests). All science grades were still substantially correlated with the domain-specific dimensions in academic secondary schools. However, the correlations between grades and the domain-specific dimensions no longer differed in a statistically significant way from the correlations between the respective grades and the GCA dimension. As expected, in the nested model the grade in German did not show a higher or lower correlation with the domain-specific dimensions than with the domain-general dimension. An exception is the correlation between the standards-oriented dimension and the grade in German, which is significantly higher than the correlation between the overarching g-dimension and the grade in German. This is not true for the mathematics grade, for which we had expected to find the same non-differing pattern. The mathematics grade correlated more strongly with the overarching g-dimension than with the domain-specific dimensions in non-academic secondary schools. As one can clearly conclude from these results, once the domain-specific abilities are controlled for *g*, the expected pattern only partly confirmed in the correlated-dimensions model disappeared.

4. Discussion

On the basis of results for the domain-specificity of verbal and mathematical achievement tests, we investigated the relation between science achievement and GCA for two prevalent assessment types: standards-oriented and literacy-oriented LSA. Structural modeling showed that just as in mathematics and verbal achievement LSA in science also measure more than domain general abilities. The relation of

Table 5
Correlations of domain-specific achievement and g/GCA with grades in German, mathematics, biology, physics, and chemistry.

	German		Mathematics		Biology		Chemistry		Physics	
	ASS	NAS	ASS	NAS	ASS	NAS	ASS	NAS	ASS	NAS
Correlated-Dimensions model										
Literacy-oriented	.34*	.14*	.39*	.18*	.36*	.30	.33*	.16*	.40*	.15*
Standards-oriented CK	.38*	.19*	.41*	.18*	.34*	.35	.34*	.20*	.45*	.17*
Standards-oriented SI	.43*	.18*	.45*	.19*	.32*	.34	.32*	.22*	.44*	.16*
GCA	.16*	.13*	.35*	.31*	.20*	.26	.25*	.22*	.25*	.16*
Nested model										
Literacy-oriented’	.31*	.09	.21*	.01	.27*	.15	.20*	.07	.30*	.11
Standards-oriented CK’	.36*	.16*	.24*	.02	.25*	.22*	.22*	.13	.36*	.13
Standards-oriented SI’	.42*	.14*	.31*	.06	.23*	.22*	.20*	.16*	.35*	.13
g	.16*	.12*	.36*	.28*	.23*	.28*	.27*	.18*	.27*	.12

Note. CK/CK’ = content knowledge; SI/SI’ = scientific inquiry; GCA = general cognitive abilities; *g* = general dimension; ASS = academic secondary schools; NAS = non-academic secondary schools.

* $p < .05$.

Table 6
Wald-test statistics on the tests of equality of correlation pairs for grades in Germn, mathematics, biology, chemistry, and physics.

Grade	German				Mathematics				Biology			
School type	ASS		NAS		ASS		NAS		ASS		NAS	
	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*
Correlated-Dimensions model												
L-OvsGCA	15.35	**<.001**	0.03	.862	0.79	.378	4.52	**<.050**	11.97	**<.001**	0.25	.615
S-O-CKvsCGA	16.54	**<.001**	1.07	.301	1.48	.223	3.21	.073	3.06	.080	1.59	.208
S-O-SI-GCA	18.89	**<.001**	0.43	.510	2.76	.097	2.83	.092	2.71	.100	1.26	.261
Nested model												
L-O'vsg	2.02	.155	0.10	.747	1.82	.177	6.99	**<.010**	0.11	.740	0.75	.385
S-O-CK'vsg	3.00	.083	0.19	.665	1.14	.285	5.12	**<.050**	0.02	.898	0.15	.704
S-O-SI'vsg	4.96	**<.050**	0.05	.823	0.13	.715	5.11	**<.050**	0.00	.996	0.19	.665

Notes. L-O = literacy-oriented; S-O = standards-oriented; GCA = general cognitive abilities; *g* = general dimension; CK/CK' = content knowledge; SI/SI' = scientific inquiry; AAS = academic secondary schools, NAS = non-academic secondary schools; Significant Wald-test results are displayed in bold.

Grade	Chemistry				Physics			
School type	ASS		NAS		ASS		NAS	
	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*
Correlated-Dimensions model								
L-OvsGCA	4.06	**<.050**	1.02	.312	16.53	**<.001**	0.02	.903
S-O-CKvsCGA	2.19	.139	0.06	.805	9.81	**<.010**	0.02	.904
S-O-SI-GCA	1.85	.174	0.00	.985	10.72	**<.010**	0.01	.920
Nested model								
L-O'vsg	0.59	.443	1.37	.242	0.11	.742	0.00	.955
S-O-CK'vsg	0.20	.654	0.20	.655	0.67	.414	0.01	.941
S-O-SI'vsg	0.46	.498	0.06	.812	0.52	.472	0.00	.981

the science variables to school grades in various subjects only partly indicated the science-specificity of the standards-oriented and literacy-oriented science LSA.

First and foremost, we showed that science achievement—as tested in LSA—is separable from GCA. Models representing science-specific dimensions were superior to the *g* model. This result is in line with earlier research on mathematical and verbal achievement (Baumert et al., 2007; Brunner, 2008; Saß et al., 2017). Since the analyses we performed on the measurement models showed that the standards-oriented science assessment was not one-dimensional, the likelihood of the *g* model best representing the data was not high. Our analyses confirmed this first indication. The correlations between the extracted science dimensions and GCA further corroborated the science specificity of the LSA dimensions. Our study design enabled us to show that this is the case for standards-oriented and literacy-oriented LSA. Thus, our findings support the notion that the educational monitoring function of science LSA goes beyond assessing a domain general ability, which means that LSA also perform its primary function in science (e.g., Koenig et al., 2008; Leutner et al., 2008; Rindermann, 2007). It was argued that the reporting of LSA was not based on domain-specific achievement (e.g., Koenig et al., 2008; Rindermann, 2007). Extending existing evidence for the domain specificity in mathematical and verbal LSA (Baumert et al., 2007; Brunner, 2008; Saß et al., 2017) to science achievement, our research corroborates the view that tests used for the three major disciplines commonly considered in LSA measure more than domain general abilities. In consequence, reported results do serve their educational monitoring function.

However, our results also indicate that the domain-specific dimensions need to be investigated in more detail. The relations to various school grades are not as pronounced as expected. The inconclusive aspects of our findings could be explained in various ways. The pattern for the school grade in German might have occurred, because the science assessment requires more reading and answering in written form, which is not the case for the GCA test. The pattern for the school grade in mathematics turned out as anticipated. While in the correlated-dimensions model the mathematics grade partially showed the expected pattern, it correlated more strongly with the overarching *g*-dimension than with the domain-specific dimensions in the nested model. This finding is consistent with the substantial relationship between quantitative reasoning and *g* (Carroll, 1993; see also Vernon, 1964).

The differing and unexpected patterns for the three science grades might derive from the multidisciplinary nature of the natural sciences. Unlike mathematical and verbal achievement, science is comprised of several disciplines that could be included with a different weight in different science LSA. Interestingly, while the relations regarding the literacy-oriented assessment and the science grades showed the expected pattern in the correlated-dimensions model in academic secondary schools, all relations with the physics grade throughout the two assessment types in academic secondary schools were as expected. This differential pattern regarding the two assessment types shows that as compared to the literacy-oriented pattern the standards-oriented LSA might be more influenced by physics than by the other two science disciplines. Thus, a future comparison of LSA could aim at content differences between literacy-oriented and standards-oriented assessments in the light of this interdisciplinarity. If different LSA put emphasis on different disciplines, they do not measure the same construct of science and the reporting of the results should consider that fact. Moreover, our robustness check on school type showed that the relations differed for academic and non-academic schools. In consequence, future studies need to prove the relations for both school types. If the complexity of the models allows for multilevel analyses, they need to be implemented as well.

Since learning materials such as school books and assessments in schools such as class tests also require reading and writing skills, our results on relations with school grades are equally relevant for science education. It could be worthwhile to investigate how these skills as well as abilities in mathematics influence knowledge acquisition in science in various learning environments and define its role in science achievement.

Interestingly, the literacy-oriented LSA correlated more strongly with the science grades (except for physics) as well as the German and mathematics grade than the standards-oriented LSA. Unlike mathematics achievement, which is clearly more closely associated with the mathematics grade and verbal achievement, which is associated with the German grade, the relation of science achievement to grades seems to be less clear. In order to acquire knowledge in science, students need a wider array of abilities from other subjects (e.g., language and mathematics skills). Another reason for the unclear picture could be that school grades in general do not only reflect objective abilities in the given subject. They also contain a students' relative position in the reference group of the classroom (e.g., Südkamp & Möller, 2009; Zeidner & Schleyer, 1998). In addition, they reflect students' personal characteristics such as motivation or conscientiousness (e.g., Krämer & Zimmermann, 2020; Ritts, Patterson, & Tubbs, 1992; Südkamp, Kaiser, & Möller, 2012). So, in disciplines for which students need multidisciplinary skills and abilities, school grades might not be the most robust external criterion to determine the specificity of the LSA. Our differential results regarding the subgroups academic secondary schools and non-academic secondary schools indicate such a subjectivity. It might be useful to consider other criteria such as ability self-concept or motivation in science and non-science subjects. Further research should examine differential patterns between the three sciences as well as the composition of the school grades in order to shed light on the weighing of the science LSA across the science disciplines. Another line of future research may explore the composition of LSA (in science) in greater detail. Rindermann and Baumeister (2015) gathered seven factors, which might explain the high correlations between GCA measures and domain-specific measures. These factors could be investigated simultaneously in a broad interdisciplinary study.

As we considered student achievement in standards-oriented and literacy-oriented assessments, our results inform stakeholders and test developers on appropriate reporting of students' achievement in a broad variety of science LSA. Results from both test types can indeed be interpreted as domain-specific outcomes that are related but not equal to a general dimension of cognitive abilities.

We investigated a LSA that was low-stakes at the student level. In Germany, low-stakes LSA (at the student level) comprise the majority of the educational monitoring system. In other countries such as the United States or Great Britain, LSA have an effect on students' school performance, students' university entrance, or on teachers' salaries (Amrein & Berliner, 2002; Emler, Zhao, Deng, Yin, & Wang, 2019). The low-stakes setting might have consequences for the relation between performance in LSA and other relevant constructs. Previous studies selectively investigated these relations, yet with inconsistent results: They found either small (e.g., Immekus & McGee, 2016) or no differences between different stakes conditions (e.g., Baumert & Demmrich, 2001), or no effects of low-stakes assessment on attitudes and motivation (e.g., Zilberberg, Finney, Marsh, & Anderson, 2014). For example, Immekus and McGee (2016) revealed that English Language Learners (ELLs) had higher test taking effort in both low- and high stakes LSA and negligibly higher importance values towards both LSA conditions. In an experimental study, Baumert and Demmrich (2001) found no effect of the study conditions informational feedback, grading and performance-contingent financial reward on the students' value of performing well, actual test performance, effort, and test motivation. Only few studies investigated methodological effects of low-stakes as compared to high-stakes testing. For instance, one study explored psychometric properties of an expectancy-value-based questionnaire administered as high- and low-stakes (Knekta & Eklöf, 2014) and found comparable properties. In our study, we focused on the domain-specificity and domain-generality of literacy-oriented and standards-oriented science LSA. A future study could divert the focus to low-stakes and high-stakes LSA and investigate whether the stake influences the domain-specificity of the assessment.

In Germany—were our study took place—the core disciplines German, mathematics, foreign language, and science are covered in both national and international assessment programs. In other countries like the USA or Great Britain, various subjects such as history or economics are included in the monitoring due to different developments of LSA (e.g., for the US see https://collegereadiness.collegeboard.org/sat-subject-tests/subjects; for Great Britain see https://www.aqa.org.uk/qualifications). Therefore, from an international point of view, comparable analyses need to be transferred to all subjects that an educational system reports on or for which the test is of great importance for stakeholders in the education system (e.g., students or teachers).

There are a number of limitations to this study that must be acknowledged. First, we based our analyses on a cross-sectional study. In most countries, the educational monitoring system comprises several age groups (e.g., for Germany, Kultusministerkonferenz [Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany, KMK], 2016; for the UK Merrell, 2017). Hence, results on domain-specificity need to be corroborated for all relevant age groups. Findings on mathematical achievement indicate that the domain specificity might be robust for science achievement as well. Second, we assessed GCA with 16 items, which were limited to fluid intelligence and a single item format. A broader operationalization of GCA in future studies would give insights into more differentiated relations between science achievement or achievement in general and GCA. Third, regarding external criteria we could only rely on a single criterion—school grades. Even though we already included school grades in various subjects, the pattern only revealed first insights into states of the domain-specific dimensions. In order to explore this issue in detail, future studies should incorporate a wide array of external criteria.

5. Conclusion

LSA in various disciplines have an impact on educational policy, on the development of curricula, and on educational practices. Hence, the conceptualization of LSA influences these practical fields as well as research in education and educators (Kind, 2013).

This study adds to previous research on adequate reporting in educational monitoring. First, we extended the investigation of *g* in education to science achievement. We found compelling evidence that science achievement also measured more than *g*. Second, we incorporated a comparison between literacy-oriented and standards-oriented science achievement, the two most prominent assessment types. It therefore seems reasonable to generalize our findings in stating that science achievement as assessed by both assessment types constitutes more than *g*. Thus, our results provide further validity evidence on LSA reporting in science achievement for various assessment types.

Acknowledgements

The data were made available by the Research Data Centre at the Institute for Educational Quality Improvement (FDZ at IQB).

Appendix A

Table A1
Correlations between general science, literacy-oriented science, standards-oriented science, and school grades in German, mathematics, biology, physics, and chemistry.

	2	3	4	5	6	7	8
General science	–	–	.27*	.27*	.34*	.26*	.26*
Literacy-oriented(1)	.84*	.80*	.24*	.25*	.34*	.24*	.23*
Standards-oriented$_{CK}$ (2)	–	.90*	.30*	.27*	.33*	.27*	.28*
Standards-oriented$_{SI}$ (3)	–	–	.30*	.29*	.32*	.27*	.27*
German (4)	–	–	–	.43*	.43*	.42*	.40*
Mathematics (5)	–	–	–	–	.47*	.56*	.56*
Biology (6)	–	–	–	–	–	.58*	.47*
Chemistry (7)	–	–	–	–	–	–	.55*
Physics (8)	–	–	–	–	–	–	–

Notes. CK = content knowledge; SI = scientific inqiry.

* $p < .05$.

Table A2
Fit indices for further models that deepen the understanding of the data.

Model	Npar	Log-likelihood	AIC	BIC	cAIC	Discrimination parameters
Global *g*-dimension	231	−61,846	124,154	125,413	125,644	Literacy-oriented = .47 Standards-oriented$_{CK}$ = .42 Standards-oriented$_{SI}$ = .49 GCA = .33 Grades = .45, .37, .43, .38, .39
Correlated-dimensions with grades	236	−61,549	123,570	124,856	125,092	Literacy-oriented = .46 Standards-oriented$_{CK}$ = .48 Standards-oriented$_{SI}$ = 51 GCA = .33 Grades = .71, .66, .72, .69, .69

Note. Npar = Number of parameters; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; cAIC = Corrected Akaike Information Criterion; CK = content knowledge, SI = scientific inquiry; GCA = general cognitive abilities. In the global *g*-dimension model all items of the literacy-oriented and standards-oriented LSA, the GCA test and all school grades load on one single *g*-dimension; the correlated-dimensions with grades model in incorporates the science dimensions of the correlated-dimensions model from Table 2 and extends it to a school grades factor that incorporates all considered school grades. Discrimination parameters in the IRT framework are the equivalent to factor loadings in factor analysis. Discriminations parameters of grades are displayed in the following order: German, mathematics, biology, chemistry, and physics. All discrimination parameter oare statistically significant on the level $p < .05$.

Table A3
Correlations of domain-specific achievement and a global school grades factor.

Construct	2	3	4
Literacy-oriented (1)	.84*	.81*	.37*
Standards-oriented$_{CK}$ (2)	–	.90*	.41*
Standards-oriented$_{SI}$ (3)	–	–	.42*
School grade (4)	–	–	–

* $p < .05$.

References

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18), 1–74. https://doi.org/10.14507/epaa.v10n18.2002.

Baumert, J., Brunner, M., Lüdtke, O., & Trautwein, U. (2007). Was messen internationale Schulleistungsstudien? Resultate kumulativer Wissenserwerbsprozesse. Eine Antwort auf Heiner Rindermann. *Psychologische Rundschau, 58*(2), 118–128. https://doi.org/10.1026/0033-3042.58.2.118.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462. https://doi.org/10.1007/BF03173192.

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Brunner, M. (2008). No g in education? *Learning and Individual Differences, 18*, 152–165. https://doi.org/10.1016/j.lindif.2007.08.005.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Carroll, J. B. (1994). Cognitive abilities: Constructing a theory from data. In D. K. Detterman (Ed.), *Theories of intelligence: 4. Current topics in human intelligence* (pp. 43–63). Norwood, NJ: Ablex.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1–22. https://doi.org/10.1037/h0046743.

Emler, T. E., Zhao, Y., Deng, J., Yin, D., & Wang, Y. (2019). Side effects of large-scale assessments in education. *ECNU Review of Education, 2*, 279–296.

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence, 62*, 138–147.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179–203.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*, 407–434.

Heitmann, P., Hecht, M., Schwanewedel, J., & Schipolowski, S. (2017). Students' argumentative writing skills in science and first language education: Commonalities and differences. *International Journal of Science Education, 36*, 3148–3170. https://doi.org/10.1080/09500693.2014.962644.

Hessisches Kultusministerium (Hrsg.). (n.d.). Bildungsstandards und Inhaltsfelder. Das neue Kerncurriculum für Hessen. Sekundarstufe I – Gymnasium. Biologie. [Educational Standards and content areas. The new core curriculum for Hesse. Lower secondary level – Gymnasium. Biology]. https://kultusministerium.hessen.de/schulsystem/bildungsstandards-kerncurricula-und-lehrplaene/kerncurricula/sekundarstufe-i/biologie.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: The Guilford Press.

Immekus, J. C., & McGee, D. (2016). The measurement invariance of the student opinion scale across English and non-English language learner students within the context of low- and high-stakes assessments. *Frontiers in Psychology, 7*(1352), 1–10. https://doi.org/10.3389/psyg.2016.01352.

Jensen, A. R. (1998). *The g factor. The science of mental ability*. Westport: Praeger.

Kampa, N. (2012). Aspekte der Validierung eines Tests zur Kompetenz in Biologie. In *Eine Studie zur Kompetenz in Biologie und ihren Teildimensionen Konzept- und Prozesswissen [Aspects of validity of a test on competence in biology. A study on competence in biology and its subdimensions content and process knowledge]*. Germany: Humboldt-Universität Berlin (unpublished dissertation).

Kampa, N., & Köller, O. (2016). German national proficiency scales in biology – Internal structure, relations to general cognitive abilities and verbal skills. *Science Education, 100*, 903–922. https://doi.org/10.1002/sce.21227.

Kind, P. M. (2013). Conceptualizing the science curriculum: 40 years of developing assessment frameworks in three large-scale assessments. *Science Education, 97*, 671–694. https://doi.org/10.1002/sce. 21070.

Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In M. von Davier, E. Gonzalez, E. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115–147). New York: Springer.

Klieme, E. (2016). TIMSS 2015 and PISA 2015. How are they related on the country level?. In *DIPF working paper*. https://pisa.dipf.de/de/pdf-ordner/Klieme_TIMSS2015andPISA2015.pdf.

KMK. (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring [global strategy of the KMK on educational monitoring]*. Bonn: KMK.

Knekta, E., & Eklöf, H. (2014). Modeling the test-taking motivation construct through investigation of psychometric properties of an expectancy-value-based questionnaire. *Journal of Psychoeducational Assessment, 33*, 662–673. https://doi.org/10.1177/0734282914551956.

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence, 36*, 153–160. https://doi.org/10.1016/j.intell.2007.03.005.

Krämer, S., & Zimmermann, F. (2020). Zum Einfluss von störendem Schülerverhalten im Unterricht auf Leistungsbeurteilungen: Explizite Einschätzungen und experimentelle Befunde [Influence of disturbing classroom behavior on teacher judgments: Explicit estimates and experimental findings]. *Zeitschrift für Pädagogische Psychologie, 34*, 99–115. https://doi.org/10.1024/1010-0652/a000250.

Lenski, A. E., Hecht, M., Penk, C., Milles, F., Mezger, M., Heitmann, P., … Pant, H. A. (2016). *IQB-Ländervergleich 2012. Skalenhandbuch zur Dokumentation der Erhebungsinstrumente [the IQB national assessment study 2012. Scales handbook of the documentation of the instruments]*. Berlin: Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen. https://doi.org/10.20386/HUB-42547.

Leutner, D., Hartig, J., & Jude, N. (2008). Measuring competencies: Introduction to concepts and questions of assessment in education. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 177–192). Cambridge: Hogrefe.

Lietz, P., & Tobin, M. (2016). The impact of large-scale assessments in education on education policy: Evidence from around the world. *Research Papers in Education, 31* (5), 499–501. https://doi.org/10.1080/02671522.2016.1225918.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the faints of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004.

McGrew, K. S., & Wendling, B. J. (2010). Cattell-Horn-Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47*, 651–675.

Merrell, C. (2017). Understanding monitoring in the United Kingdom context. In V. Scherman, R. J. Bosker, & S. J. Howie (Eds.), *Monitoring the quality in schools: Examples of feedback into systems from developed and emerging economics* (pp. 93–106). Rotterdam: Sense Publishers.

Mullis, I. V. S., & Martin, M. O. (2013). *TIMSS 2015 assessment frameworks*. Boston: TIMSS & PIRLS International Study Center.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

National Research Council. (2014). *Developing assessments for the next generation science standards*. Washington, DC: The National Academies Press.

Organisation for Economic Co-operation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2014a). PISA 2012 results: What students know and can do. In *Student performance in mathematics, reading and science*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2014b). *PISA 2012 technical report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 results (volume 1): Excellence and equity in education*. Paris: OECD Publishing. https://doi.org/10.1787/9789264266490.en10.1787/9789264266490.en.

Pant, H. A., Stanat, P., Hecht, M., Heitmann, P., Jansen, M., & Lenski, A. E. (2015). *IQB-Ländervergleich Mathematik und Naturwissenschaften 2021 (IQB-LV 2021). Version: 1 [The IQB national assessment study 2021. Competencies in mathematics and the sciences at the end of secondary level I]. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz*. https://doi.org/10.5159/IQB_LV_2012_v1.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *IQB Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I. [The IQB national assessment study 2012. Competencies in mathematics and the sciences at the end of secondary level I.]*. Münster: Waxmann.

Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*, 831–841.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*, 59–81. https://doi.org/10.1080/00461520.2016.1145550.

Rindermann, H. (2007). The g-factor of international cognitive abilitiy comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality, 21*, 667–706. https://doi.org/10.1002/per.634.

Rindermann, H., & Baumeister, A. E. E. (2015). Validating the interpretations of PISA and TIMSS tasks: A rating study. *International Journal of Testing, 15*, 1–22.

Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgments of physically attractive students: A review. *Review of Educational Research, 62*, 413–426. https://doi.org/10.3102/00346543062004413.

Saß, S., Kampa, N., & Köller, O. (2017). The interplay of g and mathematical abilities in large-scale assessments across grades. *Intelligence, 63*, 33–44. https://doi.org/10.1016/j.intell.2017.05.001.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled differences chi-square test statistic. *Psychometrika, 75*, 243–248. https://doi.org/10.1007/S11336-009-9135-Y.

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence, 46*, 156–168. https://doi.org/10.1016/j.intell.2014.05.014.

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2015). Age-related changes in the mean and covariance structure of fluid and crystallized intelligence in childhood and adolescence. *Intelligence, 48*, 15–29. https://doi.org/10.1016/j.intell.2014.10.006.

Songer, N., & Ruiz-Primo, M. A. (2012). Assessment and science education: Our essential new priority? *Journal of Research in Science Teaching, 49*, 683–690. https://doi.org/10.1002/tea.21033.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *The American Journal of Psychology, 15*, 201–293. https://doi.org/10.2307/1412107.

Stanat, S., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (2019). *IQB-Bildungstrend 2018. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich [IQB trends in student achievement 2018. The second national assessment of mathematics and science proficiencies at the end of ninth grade]*. Münster: Waxmann.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743–762. https://doi.org/10.1037/a0027627.

Südkamp, A., & Möller, J. (2009). Referenzgruppeneffekte im Simulierten Klassenraum. [reference-group-effects in a simulated classroom: Direct and indirect judgments]. *Zeitschrift für Pädagogische Psychologie, 23*, 161–174.

Undheim, J. O., & Gustafsson, J.-E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research, 22*, 149–171. https://doi.org/10.1207/s15327906mbr2202_2.

Vernon, P. E. (1964). *The structure of human abilities*. New York: John Wiley & Sons.

Wagemaker, H. (2014). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment. Background, technical issues, and methods of data analysis* (pp. 11–36). Boca Raton: CRC Press.

Wee, S. (2018). Aligning predictor-criterion bandwidths: Specific abilities as predictors of specific performance. *Journal of Intelligence, 6*, 1–14. https://doi.org/10.3390/jintelligence6030040.

Wilhelm, O. (2004). Measuring reasoning ability. In O. Wilhelm, & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 373–392). Thousand Oaks, CA: Sage Publications.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck, & M. Wilson (Hrsg.) (Eds.), *Explanatory item response models. A generalized linear and nonlinear approach* (pp. 43–74). New York: Springer.

Wu, M. (2010). Comparing the similarities and differences of PISA 2003 and TIMSS. In *OECD Education working papers, no. 32*. Paris: OECD Publishing. https://doi.org/10.1787/5km4psnm13nx-en.

Zeidner, M., & Schleyer, E. J. (1998). The Big-Fish—Little-Pond Effect for academic self-concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology, 24*, 305–329. https://doi.org/10.1006/ceps.1998.0985.

Ziegler, M., & Peikert, A. (2018). How specific abilities might throw 'g' a curve: An idea on how to capitalize on the predictive validity of specific cognitive abilities. *Journal of Intelligence, 6*(41), 1–21. https://doi.org/10.3390/jintelligence6030041.

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing, 14*, 360–384. https://doi.org/10.1080/15305058.2014.928301.