How much do students' scores in PISA reflect general intelligence and how much do they reflect specific abilities?

Artur Pokropek [1*]

Gary N. Marks[2]

Francesca Borgonovi[3]

***Corresponding Author.**

1. Educational Research Institute (IBE), Górczewska 8,

01-180 Warsaw, Poland. a.pokropek@ibe.edu.pl; artur.pokropek@gmail.com

2. Department of Sociology, Social and Political Sciences, University of Melbourne, Melbourne, Vic., Australia. g.marks@unimelb.edu.au

3. UCL Social Research Institute, University College London, London, United Kingdom. F.borgonovi@ucl.ac.uk

Keywords: *g*-factor; PISA; Raven Progressive matrix; IRT; bifactor model

How much do students' scores in PISA reflect general intelligence and how much do they reflect specific abilities?

Abstract

International Large-Scale Assessments (LSAs) allow comparisons of education systems' effectiveness in promoting student learning in specific domains, such as reading, mathematics and science. However, it has been argued that students' scores in International LSAs mostly reflect general cognitive ability (*g*). This study examines the extent to which students' scores in reading, mathematics, science and in a Raven's Progressive Matrices test reflect general ability *g* and domain-specific abilities with data from 3472 Polish students who participated in the OECD's 2009 Programme for International Student Assessment (PISA) and who were retested with the same PISA instruments, but a different item set, in 2010. Variance in students' responses to test items is explained better by a bifactor Item Response Theory (IRT) model than by the multidimensional IRT model routinely employed to scale PISA and other LSAs. The bifactor IRT model assumes that non-*g* factors (reading, math, science and Raven's test) are uncorrelated with *g* and with each other. The bifactor model generates specific ability factors with more theoretically credible relationships with criterion variables than the standard model. Further analyses of the bifactor model indicate that the domain specific factors are not reliable enough to be interpreted meaningfully. They lie somewhere between unreliable measures of domain specific abilities and nuisance factors reflecting measurement error. The finding that PISA achievement scores reflect mostly *g*, which may arise because PISA aims to test broad abilities in a variety of contexts or may be a general characteristic of LSAs and national achievement tests.

Keywords: g-factor; PISA; Raven Progressive matrix; IRT; bifactor model

Educational Impact And Implications Statement

This work uses Programme for International Student Assessment data from Poland to establish how much the achievement of secondary school students in reading, mathematics, science and in a Raven's Progressive Matrices test reflects general ability and how much it reflects domain-specific abilities. Findings indicate that a scaling model that accounts for general ability, fit data better than models routinely employed in Large Scale Assessments that ignore the role of general ability in determining individual achievement. The finding that students' responses to PISA test items reflect general ability rather than domain-specific abilities, if replicated to other countries, could have important implications for the design of large-scale assessments and the interpretation of analyses conducted using large-scale assessment data.

How much do students' scores in PISA reflect general intelligence and how much do they reflect specific abilities?

Introduction

Standardized achievement tests were developed to measure student performance in specific subject domains allowing comparisons of different groups of students, schools and jurisdictions while intelligence tests were designed to measure general aptitude and intellectual capacity. Whether the two goals overlap, and to what extent they do, is an empirical question which we address in this manuscript. Gottfredson's (Gottfredson et al., 1997, p.13) and Neisser's (Neisser et al., 1996, p.77) definitions of intelligence identify intelligence as the capacity to solve problems, understanding complex ideas and thinking abstractly while learning from experience and adapting to the context in which reasoning takes place, a capacity that, according to Hunt (2011, p.20) is "produced by an interaction between genetic potential and environmental support".

One of the most replicated findings in cognitive psychology is Spearman's identification of a general ability *g* factor that could explain between 50% and 60% of the variance in children's subject grades (Spearman, 1904; Lubinski, 2004). Although the *g* model has been criticized and alternative models offered, hundreds of studies using a variety of mental tests in different populations have also isolated *g* factors (Gustafsson & Undheim, 1996; Warne & Burningham, 2019). General ability factors isolated from different IQ tests are highly correlated (Floyd, Reynolds, Farmer, & Kranzler, 2013; Johnson, Bouchard, Krueger, McGue, & Gottesman, 2004; Johnson, Nijenhuis, & Bouchard, 2008).

Researchers have investigated the observed 'positive manifold', test-specific 'nuisance' variance, and group factors. A prominent model of human intelligence is Carroll's (1993) three strata model comprising a lower order stratum of 50 to 60 narrowly defined independent abilities, a second stratum of 8 to 10 broad independent abilities and a higher single factor of general intellectual ability 'g'. The Cattell-Horn-Carroll (CHC) factor model integrates the Carroll model with Cattell's (1963) distinction between fluid and crystallized intelligence at the middle stratum level. Fluid intelligence is independent of acquired knowledge, whereas crystallized intelligence involves acquired knowledge.

Large Scale Assessments of Student Achievement

Two large organizations are largely responsible for the design and implementation of international Large-Scale Assessments (LSA) of student achievement. The International Association for the Evaluation of Educational Achievement (IEA) administers the Progress in International Reading Literacy Study (PIRLS) of grade 4 students and the Trends in International Mathematics and Science Study (TIMSS) which monitors mathematics and science performance in grades 4 and 8. The OECD's Programme for International Student Assessment (PISA) measures performance in reading, mathematical and science literacy of 15-year-old-students. Unlike PIRLS and TIMSS, PISA is not tied to specific knowledge and skills taught at school, but aims to assess general life skills (Egelund, 2008; Schleicher, 2007). When international and national LSA reports are released, they are often followed by academic and public debate on: what they measure, their relationships with sociodemographic and educational factors, the usefulness of what they measure for students, teachers, parents and policymakers; and their unintended consequences for teachers and schools (Coburn, Hill, & Spillane, 2016; Hopfenbeck and Kjærnsli, 2016; Marsh, Roediger, Bjork, & Bjork, 2007; Sellar and Lingard, 2013; Zhao, 2020). PISA has been central to policy debates on educational reform, especially about school tracking and other forms of educational differentiation (Ertl, 2006; Grek, 2009; Breakspear, 2012; Takayama, 2008). The definition of literacy in PISA is very similar to definitions of intelligence. Literacy in reading, mathematics or science is “concerned with the capacity of students to apply knowledge and skills in key subject areas and to analyze, reason and communicate effectively as they pose, solve and interpret problems in a variety of situations” (OECD, 2007, p.16).

The usefulness of international LSAs has been questioned by studies claiming that they largely measure general cognitive ability rather than specific subject-based competencies. At the country level, mean achievement and intelligence are highly correlated (Rindermann 2007; Lynn et. al., 2012; Lynn et. al., 2009; Koenig, Frey, & Detterman, 2008). However, a high correlation at the country level does not imply that the two concepts are equivalent at the individual level (see Allik et. al. 2007 for a debate on the issue). General cognitive ability and student achievement are also highly correlated at the student level. Walberg (1984) computed an average correlation of .71 between various IQ measures and academic achievement. Kaufman et al. (2012) using structural equation modelling estimated correlations of .77 at around age 5 to above .85 at ages 16 and 17, between latent *g* and a latent academic ability factor that underlies tests of reading, math, and writing achievement. According to Zaboski, Kranzler, and Gage’s (2018) meta-analysis the correlations of *g* with basic reading, reading comprehension and basic mathematics were all above .7. The correlations with specific abilities were much lower.

The general model used in LSAs for student abilities in specific subject domains is quite different to the models of human intelligence. Psychometricians working on LSAs typically specify latent ability factors—reading, mathematics and science in PISA; mathematics and science in TIMSS— as distinct and correlated constructs. Multidimensional models generate students’ test scores which are then analyzed and reported. The underlying model is referred to as the standard model or multidimensional model. The multidimensional model assumes that each assessment domain corresponds to a single latent factor which fully represents capability in the specific domain. These latent factors are correlated with each other. The general ability factor *g* is assumed to be irrelevant to students' test scores. However, the constructs isolated in the standard model contain a considerable amount of variance attributable to *g* (Brunner, 2008).

The Bifactor Model

An alternative model – the bifactor model (also called the nested-factor model) - effectively reconciles educational assessment research and intelligence research. It specifies a general ability *g* factor and domain specific ability factors underlying variation in student performance in achievement tests (Gustafsson & Balke, 1993) or in intelligence tests (Gignac & Watkins, 2013). In the bifactor model, latent constructs for general and specific abilities are specified as uncorrelated with, or orthogonal to, each other. Unlike the three stratum or CHC models there is no hierarchy in the bifactor model: test items load directly on both general and specific latent ability factors. Figure 1 illustrates four models describing potential relations between general ability *g* and domain-specific abilities and how these relate to response patterns to specific test questions administered in the context of achievement tests.

FIGURE 1

Generally, the bifactor model tends to fit the data from intelligence tests better than models with a hierarchical structure (Cucina & Byle, 2017). In the bifactor model, specific factors typically explain less of the total variance than the specific factors in the CHC and multidimensional models (Jensen & Weng, 1994; Eid, Krumm, Koch, & Schulze, 2018). The specific factors in the bifactor model are 'purer' uncorrelated with general ability. Although it makes little difference in practice if multidimensional hierarchical or bifactor models are used when the focus of the analysis is on *g*, the use of multidimensional hierarchical or bifactor models is crucial when the interest lies in the second level factors as is the case for LSAs (Beaujean, Parkin, & Parker, 2014).

Brunner (2008) compared a two-dimensional standard model and a bifactor model using four mathematics scales and three reading scales from the German PISA 2000 study together with data from a cognitive ability test. The bifactor model exhibited slightly better fit indices. He found that *g* explained 40% of the variance in mathematical ability and 49% of the variance in verbal ability. The average amount of variance attributable to domain-specific abilities was much lower: 8% for mathematical ability and 17% for verbal ability. Baumert, Lüdtke, Trautwein and Brunner (2009) analyzing the same data compared the *g*-factor or Spearman model comprising only one latent factor (*g*) and a bifactor model comprising *g* and specific latent mathematical and verbal factors. The bifactor model provided a much better fit. Baumert et al. (2009, p. 173) concluded that general ability is a key determinant in the acquisition of knowledge and skills at school, and domain specific abilities make an incremental contribution to performance, above and beyond *g* but did not compare the variances accounted for by general and specific abilities. Baumert et al. (2009, p.165) emphasized that “the outcomes of schooling can and must be conceptually distinguished from the intelligence construct”. Analyzing TIMSS rather than PISA data, Saß, Kampa and Koller (2017) compared a two factor correlated model, comprising correlated latent factors for *g* and math, and a bifactor model with uncorrelated latent factors for *g* and math, for German students in grades 5, 9 and 13. Their study concluded that LSAs test mathematical ability beyond *g*.

Although it is generally agreed that LSAs measure more than just *g*, LSAs typically estimate students’ scores in different domains from multidimensional models without a general ability factor. In PISA, the reading, math and science factors isolated from the multidimensional model are highly correlated. Bond and Fox (2001) reported intercorrelations of .82 between mathematics and reading, .89 between science and reading and .85 between mathematics and science. Cromley (2009) presented correlations around .8 between reading and science in PISA 2003 and 2006 for individual countries. These correlations are higher than the correlations between simpler measures of domain scores (Marks, 2016). The very high interdomain correlations from the multidimensional model undermine the assumption that the PISA domains represent largely independent learning domains with unique influences, such as, for example students' reading habits for reading, the qualifications of mathematics teachers for mathematics and a school's science resources for science. The obvious explanation for the high interdomain correlations is that the domains incorporate substantial general cognitive ability components. A *g* factor is required to remove the contamination of domain specific abilities with general ability. A meta-analysis of 50 studies that used bifactor models published in psychopathology, personality, and assessment journals concluded that modelled variance is overwhelmingly due to a single general latent variable rather than to the specific factors often emphasized by researchers (Rodriguez, Reise, & Haviland, 2016).

The studies cited above on the latent structure of LSAs clearly identified general ability factors in addition to domain specific abilities. However, these findings did not change practices within the psychometric community involved in the construction and scaling of LSAs, nor in the education academic research and policy communities that analyze and interpret LSA data for research and to inform policy making (e.g., Keller, Preckel & Brunner 2020; Borgonovi & Pokropek 2019; Deng, & Gopinathan, 2016; Jakubowski & Pokropek 2015). Measurement frameworks in education do not consider the relevance of *g*.

Brunner (2008) notes that analyses of students' characteristics and domain-specific ability scores reported in LSAs, meta-analyses, and reviews have relied almost entirely on the standard model of domain-specific abilities. He speculates that this literature would be rather different if a bifactor conceptualization of student achievement predominated in educational research. Saß, Kampa, & Koller (2017) advise researchers to carefully choose between the two-factor correlated and the bifactor models because the math factor holds a considerable amount of *g* in the two-factor correlated model but none in the bifactor model. The choice is also guided by the plausibility of the relationships the specific factors have with covariates.

Relationships of covariates with the specific ability domains differ between the multidimensional and bifactor models. For instance, in Brunners's (2008) standard multidimensional model, socioeconomic status correlated at around .35 with both mathematical and verbal latent factors. In contrast, the corresponding correlations were much lower in the bifactor model ($r = .05$ and $r = .09$) and socioeconomic status correlated more strongly with the general *g* factor ($r = .35$). The same pattern was found for books in the home, satisfaction with school and educational aspirations. Baumert et al. (2009) also found that socioeconomic background more strongly correlated with *g* than with specific abilities. Saß, Kampa, & Koller (2017) found SES correlated slightly more with *g* ($r = .24$ for 9th grade) than with the specific math factor ($r = .15$ for 9th grade). In grade 13, the correlations with both g and the math factor were not statistically significant. Brunner (2008) found that grades in German and mathematics had more distinct relationships with math and verbal factors in the bifactor model than in the multidimensional model. In the bifactor model, girls had much higher scores than boys on the reading factor and lower scores on the math factor and gender differences on *g* were much smaller than on either domain (Baumert, 2009). Saß, Kampa, & Koller (2017) found larger gender differences favoring boys on the mathematics factor than on *g* in grades 5 and 9, but not in grade 13.

The release of international LSAs like PISA and TIMSS is followed by international and national reports and later by academic journal articles based on analyses of publicly released data. These analyses often examine associations between demographic, socioeconomic, school and attitudinal factors, and student performance in a single domain. Given that student scores generated from multidimensional models (plausible values) in LSAs are highly inter-correlated, statistical relationships between students' domain scores and covariates tailored for a particular domain—enjoyment of and frequency of reading, atmosphere in math classes, time spent in science classes, teachers' qualifications in math or science— should be evaluated in light of the fact that the domain measures are not pure measures of reading, mathematics or science; they comprise substantial common variance, most likely *g*. So, the generation of student scores from different statistical models is not an arcane statistical exercise but has consequences for the nature and strength of relationships between students' scores and putative influences and their interpretation by researchers and policymakers.

Purpose of the Present Study

Previous studies comparing the bifactor and standard models using PISA data did not analyze the actual responses provided by individual students. Brunner's (2008) analysis used 4 mathematical subscales, 3 verbal subscales and 2 cognitive ability scales. Similarly, Baumert (2009) analyzed reading and mathematics PISA subscales. Analyses of subscales assume that individual items have the same loadings for *g* and for the subscales, an assumption that is unlikely to hold for all items. In contrast to most previous studies, we use item-level information rather than subdimensions of aggregated items. Analysis of individual-items responses considerably increases statistical power producing more precise estimates, which is important for small effects. In addition, organizations that generate student scores using multidimensional models analyze individual items.

Furthermore, the specific ability factors isolated in bifactor studies developed using data from LSAs and cited above are limited to math and verbal literacy. No study has included science, and all have analyzed data from German students. This is the first study replicating the PISA's multidimensional model supplemented by Raven' ability test scores. No compromise or simplifications of the PISA measures were applied. Respondents sat the full PISA test twice.

Finally, no previous study has explicitly compared the multidimensional (non *g*) models used in a LSAs with the corresponding bifactor and hierarchical multidimensional models.

Therefore, the objectives of this study are:

1. To compare various models of students' responses to the PISA items with a particular focus on the standard multidimensional model and the bifactor model.
2. To assess the strength and reliabilities of the general and specific ability factors isolated from the bifactor model.
3. To examine if relationships of criterion variables with the specific ability factors isolated from the multidimensional and bifactor models are consistent with what would be predicted by theory.

This study analyses PISA data from Poland together with a measure of fluid intelligence to estimate the extent to which students' PISA scores reflect general ability (*g*) and specific educational abilities in reading, math and science. The superiority of bifactor model found in the contributions of Brunner (2008) and Baumert et al. (2009) might reflect the inclusion of specific measures of cognitive ability. The dominance of *g* may be less important when only the three sets of PISA test items are included. Therefore, the study further compares the multidimensional and bifactor models for just the PISA items, not including the Raven's ability test items.

Previous studies relied mostly on model fit for comparisons of different latent models. Reliance on comparisons of model fit has been criticized for being overly simplistic (e.g., Berge and Sočan 2004; Bentler 2009). For this study comparisons of model fit are supplemented with statistical indices derived mainly from the bifactor model, to identify the sources of the common variance in students' responses to test items (Reise, Bonifay, and Haviland 2018; Gignac & Kretzschmar 2017). We choose the Explained Common Variance (*ECV*) indices (Reise, Bonifay, & Haviland, 2018), the Omega bifactor model-based reliability indices (Raykov 1997; Reise, Bonifay, & Haviland, 2013; Reise at al. 2013), and Haberman's (2008) Proportional Reduction in Mean Squared Error (*PRMSE*). These indices provide more detailed analyses of the latent structures and assess the extent that domain specific factors measure domain specific abilities, or are best understood as nuisance factors reflecting imperfect measurement of general ability (Reise, Moore & Haviland, 2010). General fit measures inform only about the overall fit of the models, while the other indices allow assessment of the extent that each group of items is related with the different latent factors providing indications of misspecifications of the latent structure. Moreover, they estimate the reliabilities of the dimensions generated from different latent structures and provide additional information that can aid the understanding of variation in students' responses to individual test items.

Materials and Methods

Data

PISA is a triennial large-scale standardized assessment conducted since 2000 and targeting the schooled population of children aged between 15 and 3 months and 16 and 2 months. Each PISA cycle assesses three core domains (reading, mathematics and science). Students take a two-hour test and are then asked to complete the student questionnaire. PISA is a low stakes test because test results do not have any consequences for participants. It is a high stakes test for senior education bureaucrats because the performance of students from different countries and jurisdictions are publicly compared.

The core PISA instruments are developed, validated and administered following strict technical standards defined internationally which guarantees comparability (OECD, 2014). PISA's national options allow countries to use additional instruments and to administer the core international instruments to additional groups of students. In 2009, Polish students aged 16 or older from grade 10, the first grade in Polish upper-secondary schools, were included in the Polish national PISA option.

The same protocols and procedures used in the main PISA study were implemented for the Polish extension. The major exception was sampling. In the Polish extension of PISA, one class was selected at random from each school. By contrast, in the standard PISA sample students are selected at random within each selected school. Sampling intact classes greatly facilitates data collection but reduces sample efficiency. In contrast to the core PISA sample, the target population was defined by grade not age. Grade-based sampling is the approach used in PIRLS and TIMSS. As long as schools, and classes within schools, are sampled randomly, a classroom-based sampling strategy does not introduce systematic biases. The PISA 2009 Polish national extension formed the basis of the *From School to Work* panel study (http://www.fs2w.ifispan.waw.pl/).

The target population of the *From School to Work* study was grade 10 students attending any type of upper-secondary school in Poland. Students with certified disabilities were excluded.

In the first stage of the stratified two stage sampling procedure, schools were divided into four strata according to school-type: 100 general high-schools, 6 professional-oriented high-schools, 54 vocational secondary schools and 40 basic vocational schools. Within each stratum, schools were randomly selected with probabilities proportional to the number of grade ten classes in the school. In the second stage, one grade ten class was randomly selected in each school.

The first wave of the study was conducted in March 2009 comprising 4 951 students. Participating students completed the standard PISA 2009 instruments: the three achievement tests and the background questionnaire. Six months later, in October 2009, a second wave was conducted comprising 4 041 students; attrition was due largely to refusal or because students changed schools. Students had just begun grade 11 and were administered the Raven's (2003) Progressive Matrices test. The third wave was conducted six months later (April 2010) comprising 3 989 students with a second PISA assessment. A total of 3 472 students took part in all three waves. All students completed the tests and questionnaires in classrooms. Students completed the instruments individually, but were supervised by teachers and interviewers in accordance with PISA protocols (for more details see OECD 2009).

Measures

Academic achievement

The PISA test is administered using a rotation design: students are assigned test booklets containing only a subset of the full testing material that was developed. This is known as an incomplete balanced matrix design; each student answers a sample of test items. The item pool consisted of both multiple-choice and constructed response questions. The items varied by domain, format and difficulty (OECD 2009). Test items were organized into clusters of subject-specific items, each designed to take around 30 minutes to complete. The clusters were organized into 13 booklets, each booklet contained four clusters and each cluster was paired at least once with every other cluster. Students were administered at random one of the test booklets, and booklets were designed to contain four clusters of testing material that were rotated across booklets such that each cluster was paired at least once with each other booklet and each cluster was administered in different positions of the test (start, middle and end of the test) (OECD 2012: 29-32). Since the booklets were distributed completely at random, estimates of abilities are unbiased (for details see OECD 2012: 29-32).

Polish translations of the PISA instrument were used to measure reading achievement (105 test items), mathematics achievement (37 items), and science achievement (49 items). In 2009, reading was the major domain, hence there were far more reading than math or science items. For the Polish national option, item clusters occupied each of four possible positions in the booklet: start, early middle, late middle and end of the test. In the first wave (2009) students were randomly assigned one of the 13 booklets. For wave three they were randomly assigned one of the 12 booklets remaining after excluding the booklet they took in wave one.

Raven's Standard Progressive Matrices

The Polish adaptation of the Raven's Standard Progressive Matrices was used (Jaworowska, Szustrowa, & Raven,2000). The Raven’s test is a 60-item paper and pencil multiple choice test of non-verbal reasoning ability (Raven, 2003). Items consist of figures missing a piece. Test subjects are asked to select the correct missing piece among six or eight alternatives to complete the figure. The Raven's test shares approximately 50% of its variance with *g* (Gignac, 2015).

Sociodemographic Measures

Students participating in the study were administered the standard PISA 2009 background questionnaire. Students were asked to report the educational attainments and occupations of their parents, and respond to items on educational resources, cultural resources and the presence or numbers of consumer durables in the home. This information was used to create a composite index of socio-economic status, the PISA Index of Education, Social and Cultural Status (ESCS) which has been widely used in the policy and academic literatures (see OECD, 2012; Avvisati, 2020 for an extensive review). The index was standardized to have a mean of zero and a standard deviation of 1, across OECD countries (for more details on the index and its construction, see OECD, 2009). Other data from the PISA student questionnaire used in this study were students' reports on their expected educational attainment and their attitudes towards reading and school. Additional questions administered specifically for this study were: students' grades in mathematics, Polish and biology. Grades were assigned by teachers following guidelines from the Ministry of Education using a 1-to-6-point grading system.

Principals or designates of sampled schools were asked to complete a paper and pencil questionnaire on the school. From this information three measures were constructed on the total class hours per week that grade 10 students typically took in Polish and other humanities, science and mathematics.

Analytical strategy

The four psychometric models described in figure 1 are compared. Each model assumes a different latent structure to account for the variation in students' responses to the test items.

PISA test scores were measured twice and each model was estimated twice, first with 2009 PISA data and then with 2010 PISA data. Since, students had experienced a full year of schooling between the two PISA rounds of testing, the importance of domain specific latent factors may be larger for the 2010 data compared to the 2009 data. Differences between analyses of the two data sets may reveal the effects of an additional year of domain specific knowledge acquisition on the latent structure.

Models

Model 1: One-dimensional Item response Theory (IRT) model. The first model is a Spearman type model which specifies that all PISA and Raven's items load on one common latent factor, *g*. Student responses are directly related to the underlying unidimensional ability factor reflecting general intelligence. This model assumes that specific abilities do not contribute to explaining the variation in students' PISA scores and do not increase the probabilities that students' respond correctly to the items. This model is unlikely to fit the data well; it should be considered the departure model because it is the least constrained.

Model 2: Four-dimensional IRT model. This model assumes that four different but correlated latent traits best describe test takers' patterns of responses to the three sets of PISA items and the Raven's items. The model specifies that general ability is not necessary to describe students' responses and the factors are not independent. This model resembles the standard model used in LSAs (excluding the Raven's test).

Model 3: Higher order IRT model. This model specifies four orthogonal (uncorrelated) latent traits in reading, mathematics, science and the Raven's that correlate with, or load on, the higher order general cognitive ability factor. The higher-order IRT model described in model 3 implies full mediation: the association between the higher-order factor *g* and the observed variables $m_1...m_n$; $r_1...r_n$; $s_1...s_n$; $i_1...i_n$ are assumed to be fully mediated by the lower-order factors *Math, Read, Sci* and *Raven* (Yung, Thissen, & McLeod 1999). Model 3 resembles the Cattell–Horn–Carroll hierarchical model of intelligence in which Raven's represents fluid intelligence

(Raven, 2003 p 73). The three PISA domains of reading, math and science represent crystallized intelligence.

Model 4: Bifactor IRT model. In the bifactor model, the non-*g* factors are uncorrelated with *g* and with each other, and they comprise only specific factor variance; they are “pure” representations of the hypothesized specific abilities net of the general factor. Each of these factors accounts for some of the variance in item responses, not accounted for by the general cognitive ability factor (Reise, Bonifay & Haviland, 2013; 2018; Rodriguez, Reise & Haviland 2016). Domain specific factors may explain differential student responses between domains and why some students do not perform as well in one domain as in do in another domain, for example gender differences in math and reading which cannot logically be explained by general ability. Alternatively, they are *nuisance* factors, reflecting the imperfections of different sets of measurement instruments to adequately capture general ability (Reise, et al., 2013).

Analysis

In the first part of the paper, the fit of the four models to the observed data described in figure 1 are compared using a variety of appropriate fit measures (Hu and Bentler, 1999). In particular, we report Log-likelihood, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The Comparative fit index (CFI), Tucker-Lewis index (TLI), Root Mean Square Error of Approximation (RMSEA), or chi-square indices are not presented because PISA uses an incomplete balanced matrix design in PISA for the cognitive tests (for details see OECD 2012: 29-32). The resulting large fraction of missing data for cognitive items is inappropriate for these summary indices (Agresti, 2010). This is followed by estimating the correlations of the latent ability factors from the multidimensional (four-factor) model.

MODEL INDICES

The second part of the analysis examines the sources of common variance and their reliabilities informed by several indices: the Explained Common Variance (*ECV*) index (Reise, Bonifay, & Haviland, 2018) and the *Omega* bifactor model-based reliability indices (Raykov 1997; Reise, Bonifay, & Haviland, 2013; Reise at al. 2013). In addition, Haberman's (2008) Proportional Reduction in Mean Squared Error (*PRMSE*) statistic is included which is based on subscale scores, not the bifactor model. *PRMSE* complements *ECV* and other model-based reliability indices. The indices indicate if the domain specific factors identified in the bifactor model contain enough reliable information to be interpreted substantively or as measurement error produced from scaling. The formulas for the indices are in the online appendix.

ECVGen is the common variance explained by the general factor divided by the total common variance. *ECVGen* indicates the relative "strength" of the general factor. It is "high whenever there is little common variance beyond the variance captured by a general trait, regardless of the size of the item loadings estimated considering a single general trait" (Reise et. al. 2013, p. 11). *ECV* values on the general factor are above 0.6 (Reise, et al., 2013) or above 0.7 (Rodriguez, Reise, and Haviland, 2016) are considered high and indicating the strong dominance of a general factor – that is unidimensionality. This index could be also computed for specific factors - *ECVSp* indicating the proportion of the common variance each specific factor accounts for.

Coefficient Omega (McDonald 1999) is a factor analytic model-based reliability estimate. There are two *Omega* indices calculated differently for general and specific factors. *Omega* indicates of how much of the variance in the observed total score can be attributed to all modeled common factors, that is all factors related to a set of items. For the general factor, *Omega* is calculated from all items. For the reading, math, science and Raven's factors *Omega* is calculated only with the items belonging to the respective domain.

The *OmegaH* indices indicate how much reliable variance of the total scores can be attributed to each factor (Reise et. al., 2013). For the general factor, the higher *OmegaH* is, the more the general factor is the dominant source of systematic variation. If *OmegaH* is high (> .80), the factor structure can be considered unidimensional because the bulk of the reliable variance is due to a single common factor. *OmegaH* for subscales is the proportion of subscale score variance attributable to a group factor, after removing the reliable variance due to the general factor (Rodriguez, Reise, and Haviland, 2016).

The ratio of *Omega* to *OmegaH* quantifies how much of the reliable variance in total scores is accounted for by the general factor *g* compared to the specific factors. *Omega* and *OmegaH* and their ratios are computed for each of the five orthogonal factors. For the three PISA domains, the *Omega* ratios indicate the extent domain specific scores reflect general and specific abilities. If the ratio for a subscale is low, most of the reliable variance of the subscale scores can be attributed to the general factor. If the ratio is high there is substantial reliable and unique sub-scale variance.

PRMSE indicates the relative importance of specific factors over the general factor in explaining variability in responses to test items (Haberman 2008). The *PRMSE* ratio indicates the extent to which separate scaling increases or decreases the amount of information conveyed in the scale. If the *PRMSE* ratio for a specific subscale is greater than 1.0, the corresponding factor is considered to add information beyond that provided by the general factor. If the *PRMSE* ratio is less than 1.0, the addition of specific factors does not provide additional information.

At the conceptual level, *PRMSE* is like the *Omega* coefficients. However, *PRMSE* is computed on observed scores and does not assume the factors are orthogonal or an underlying bifactor model (Haberman, Sinharay & Puhan, 2008). *PRMSE* can therefore be considered as an additional robustness check.

Predictive Validity with Criterion Variables

In the third and final part of the paper, the predictive validity of the four-dimensional and bifactor models are examined by correlating the domain specific factors with criterion variables: grades in language of instruction, mathematics and biology; enjoyment of reading; learning time in the three subjects; gender and socio-economic status.

If the bifactor domain specific factors are substantively important and can be considered as “pure” representations of the specific hypothesized abilities (net of the general factor), they should exhibit theoretically plausible correlations with domain specific criterion variables. This expectation is guided by the reasonable assumption that spending time studying a subject, or enjoyment of that subject are associated with knowledge and abilities in that subject, net of general cognitive ability. Furthermore, higher grades in one subject should relate to that subject’s latent factor rather than other subjects’ latent factors. Therefore, the latent reading factor should correlate with learning time in humanities, grades in humanities and enjoyment of reading. Similarly, the latent math factor should correlate with learning time in math and grades in math, and the latent science factor should correlate with learning time in science and grades in science.

The PISA ESCS index is understood to reflect parents’ attitudes to education and their involvement with their children’s education, their financial, cultural and social resources that facilitate their child’s performance at school (Avvisati, 2020; OECD, 2013, p.2). Competing theories lead to alternative hypotheses of the relative strength of the association between socioeconomic status and the specific factors and the general ability factor estimated in the multidimensional and bifactor models.

According to theories of cultural reproduction and social stratification, socio-economic status shapes achievement by shaping the extent to which individuals have access to educational, material and cultural resources, through quality teaching and learning opportunities, through access to high quality information on how best to acquire knowledge (Buchmann, 2002). These factors are unlikely to influence general cognitive ability but influence specific abilities among individuals with similar underlying general abilities. To the extent that economic and cultural resources explain socio-economic differences in achievement, ESCS should correlate more strongly with the specific PISA ability factors than with general ability. Furthermore, it should correlate more strongly with the reading factor than the science or math factors, since ESCS includes measures of the number of ‘books in the home’, and the presence of ‘classic literature’ and ‘books of poetry’ in the family home.

An alternative explanation for the relationship between student achievement and socio-economic status is parents’ educational and socioeconomic attainments relate to their cognitive abilities which are transmitted either genetically or through early parental investments to their children and children’s cognitive abilities influence their performance in achievement tests. According to this explanation, ESCS would be more strongly related to general cognitive ability than the specific ability factors. Prior analyses of correlations between socio-economic condition and general cognitive ability and specific ability factors estimated with a bifactor model are in line with this expectation (Baumert et al., 2009; Saß, Kampa, & Koller, 2017).

In addition, gender differences in achievement evident in PISA and in other achievement studies (see Stoet & Geary, 2013)—girls perform better than boys in reading but vice versa for math—should be reflected in the relationships between gender and the specific ability factors. In contrast, there should be no gender difference in general cognitive ability (Halpern, 2012). The validity of the bifactor model would be undermined if correlations with criterion variables were not consistent at the two time points, or there were many correlations contrary to theoretical expectations, such as math grades, math learning time correlated with the reading factor, and being female correlated positively with math but negatively with reading.

Parameter estimates were obtained with Mplus version 7.4 (Muthén and Muthén, 1998-2017) using models for binary data (often referred to as Item Response Models). In all analyses, partially correct responses (partial credits) were designated as correct. We employed full maximum likelihood (FIML) estimation and robust standard errors to account for the multistage sampling in PISA. FIML is the most appropriate method for data missing completely at random (MCAR). This is the case in our study: as indicated previously, test booklets were distributed completely at random to participating students. Robust standard errors of correlation coefficients are as accurate as replication weights often employed in analyses of LSAs and are much more easily implemented. Replication methods should be used whenever estimates involve statistics without known distributions like percentiles or interaction effects (Efron 1982; Kolenikov, 2010) which is not applicable here.

We used a two-stage estimation strategy. In the first stage measurement models were estimated. Fit statistics from this stage were used to compare how well the data fit the latent structures illustrated in Figure 1. In the second stage, we added a saturated structural model, which allowed us to estimate the correlation coefficients between latent constructs and criterion variables. This structural model was estimated with measurement model parameters fixed at the values estimated in the first step. This procedure reflects the general approach of scaling employed in LSAs data (e.g., OECD 2012; Martin, Mullis, & Hooper 2017) without the additional step of generating plausible values. Here the parameters of interest are observed directly form the structural part of the model (see Von Davier, Gonzalez, & Mislevy, 2009). Full information maximum likelihood is used to estimate the parameters of the measurement model since it is the most appropriate method for data missing completely at random (MCAR). This is the case in our study: as indicated previously, test booklets were distributed completely at random to participating students.

Missing data in the student questionnaire, was low. There were no missing values for gender; 2% of the participants did not information on language and mathematics marks; 16% did not have information on marks in biology grades (in some schools biology was not obligatory at this stage and many students left this question unanswered as a result); less than 1% did not have information on ESCS and on enjoyment of reading; and 4% did not have information on the learning time variables. Maximum Likelihood handles missing data for covariates by analyzing only non-missing data to estimate the set of parameters with the largest likelihood. It produces unbiased estimates with data missing at random (MAR) (Graham; 2012).

For the *PRMSE* the imputation procedure used was “imputations by chained equations” (Royston, 2004) which also accommodates PISA’s rotational design (OECD 2012). The imputation model was based on all responses to items. *PRMSE* indices were calculated on the imputed dataset.

Results

Comparisons of alternative latent structures

Table 1 presents the correlations between latent constructs according to the multidimensional model (Model 2 in figure 1), the model usually employed in LSAs and which assumes correlated constructs. The lower-diagonal correlations are from the 2010 data and the upper-diagonal correlations are from the 2009 data.

The average correlation between the three PISA factors in the multidimensional model is .86. The latent correlations between the PISA factors in the 2009 and 2010 data are virtually identical. The correlations between pairs of corresponding domain specific factors in 2009 and 2010 are very high: .90 for math, .88 for reading, and .87 for science. The very strong correlations indicate that although overall levels of achievement increased over the intervening 12 months, the relative positions of students were largely unchanged.

The average correlation of the PISA factors in 2009 and 2010 with the Raven's factor was the same ($r = .73$). The Raven factor is more highly correlated with the mathematics factor than with the reading and science factors.

TABLE 1

The high inter-correlations ($.82 < r < .90$) of the three PISA factors in the multidimensional model are comparable to the correlations of the PISA domain specific abilities referred to in the literature review. A common general ability latent factor is likely to account for these very high correlations between learning domains which are purported to be substantively independent.

Table 2 presents common fit statistics for the four models summarized in figure 1. Item loadings are presented in figures 2 and 3. Table 2 reports findings for analyses performed using data from PISA 2009 and 2010. Results show that the bifactor model (figure 1: model 4) fits the data best. In both sets of analyses, model 4 exhibits the least negative log likelihood ratios, the highest scaling correction factor, and the lowest Akaike, Bayesian and sample-size adjusted BIC fit measures. Interestingly, the multidimensional model (model 2) does not provide a substantially better fit than the simple Spearman type model (model 1).

TABLE 2

Item Loadings on Factors

Figures 2 and 3 present values of standardized loadings which indicate the strength of the associations between items and factors. The loadings on the specific factors are at positions:

- 1 to 48 for the Raven's items on the Raven's factor,
- 49 to 154 for the reading items on the reading factor,
- 155 to 204 for the science items on the science factor,
- 205 to 242 for the math items on the math factor.

The loadings on the general factor are at positions:

- 243 to 290 for the Raven's items,
- 291 to 396 for the reading items,
- 397 to 446 for the science items,
- 447 to 484 for the math items.

For both the 2009 and 2010 items the loadings on the specific PISA factors are smaller than for the general factor. For the Raven's items, the loadings on the specific Raven's factor are comparable with their loadings on the general factor indicating that the Raven's test generates a latent factor independent of the general factor. The average item loading on the general factor and Raven's factors is around 0.5 with standard deviations around 0.1 In contrast, average item loadings on the specific PISA factors are much lower: around 0.2 (sd = 0.1). For the reading, math, and science factors the loadings range from below zero to 0.7 (see Table 3). Virtually all negative loadings are not statistically significant.

TABLE 3

FIGURE 2

FIGURE 3

Exploring dimensionality: domain specificity and general factor

Table 3 reports *ECV*, *Omegas* and *PRMSE* results from the bifactor model. Having established that the bifactor model is the best fitting model, the consequent issue is: are the four domain specific factors substantively meaningful or are best described as nuisance factors reflecting little more than test format and other systematic, but minor, sources of variation?

ECV results for both 2009 and 2010 data indicate that the general factor *g* explains around 70% of the common variance while the remaining part of the common variance is explained by the specific domains. The *ECV* for *g* was 0.73 for both 2009 and 2010. The *ECV* indices for the four orthogonal factors range from 3% for math and science in 2010 and mathematics in 2009 to 10-12% for reading and 12% for Raven's. These percentages are remarkedly consistent across the two time points. Math which has a distinctive cumulative curriculum accounts for only 3% in both years. It appears that variation in math is subsumed in general ability and to a lesser extent Raven's. Reading, which in secondary-school is not formally taught, appears to be more distinctive than math or science.

TABLE 4

The *Omega* coefficients reported in Table 4 confirm that most of the common variance across items can be explained by the general factor *g* while the other factors account only for a small portion of the reliability. The *Omega* coefficients indicate that the latent factors are reliable but much of the reliability of the non-*g* factor is attributable to the general ability factor. The *OmegaH* statistics suggest that the reliabilities of the non-*g* factors are not acceptable as independent factors, with the possible exception of the Raven's factor.

The *Omega* ratios are between 0.05 and 0.17 for the three PISA domains confirming that only a small amount of reliable information captured by the PISA items is domain specific. These results indicate that the PISA test items do not form distinct independent factors corresponding to their specific domain but are mainly subsumed by the general *g* factor. The only factor that may be considered independent of *g* and somewhat reliable is Raven's. For both years, around 50% of the reliable variance in Raven's test scores can be attributed to the Raven's factor. The Raven's factor is more distinctive that the three hypothesized PISA domains.

The *PRMSE* estimates are consistent with the *ECV* and *Omega* indices. The 2009 results in Table 3 show that the independent math factor reduces measurement error for the mathematics items by only 3% (final column) compared to just the general factor. The reading and science factors increase the informational content (or reduce measurement error) of the test items by about 10%. The Raven's factor increases the informational content by around 50%.

To ensure that findings on the primacy of the general factor *g* do not depend on the inclusion of the 60 Raven's items, the bi-factor model was re-estimated without the Raven's items. With only the PISA items, the *g*-factor accounts for even more of common variance. The *ECV* for the general factor was 0.76 in 2009 and 0.81 in 2010. So, the dominance of general factor and the much weaker specific factors is not because of the inclusion of the Raven's items in the model. Consistent with the analyses that included the Raven's items, the reading items are most distinctive accounting for 14% of the common variance in 2009 and 9% in 2010. The *ECVs* for math and science are very small at around 0.04. *Omega* statistics indicate that all three specific ability factors have very low reliabilities. At the same time, the *PRMSE* indices suggest that the three subscales add some information. The finding that in the absence of the Raven's items the general factor is even stronger, reiterates the conclusion that students' responses to the test items largely reflect general ability.

Correlations between latent factors and criterion variables

Table 4 presents correlations between the latent factors isolated in 2009 and 2010 from the multidimensional (four-dimensional) and bifactor models with criterion variables. In the four-dimensional model, the correlations with criterion variables were often statistically significant and mostly consistent across the two years. This was because nearly all the reliable variance in the reading, math, and science factors was due to the general ability factor, not the specific abilities they were supposed to represent. Several correlations were contrary to theoretical expectations: the high correlation of grades in mathematics with the reading factor; the positive correlations of enjoyment of reading with the math and Raven's factors (although lower than for reading and science); and the positive correlation of language grades with the science factor. These anomalous correlations most likely reflect specific factors having large *g* components.

For the specific ability factors in the bifactor model, several of the correlations between them and the criterion variables conform to theoretical expectations assuming they are pure measures of specific abilities. There are positive correlations of the math factor with math grades. Being female is positively correlated with the reading factor and negatively correlated with the math and science factors in both years. Gender differences on the specific factors tend to be larger in the bifactor model than in the four-dimensional model.

There are, however, several inconsistencies across years in the correlations of the specific ability factors from the bifactor model with criterion variables. Grades in humanities, learning time in humanities and being female are positively correlated with the reading factor in 2009 but not in 2010. Language grades are negatively associated with the math factor in 2010 but not in 2009. One explanation for these inconsistent results is that the reading, math, and science factors are rather unreliable measures of the corresponding specific abilities. So, they are somewhere between specific ability factors and nuisance factors.

Expectations that ESCS would be more strongly correlated with the specific ability factors, especially reading, were not realized. According to the multidimensional model the association between each PISA latent domain specific factor and the PISA measure of socioeconomic status (ESCS index) is 0.08 for reading and science and 0.12 for mathematics. In contrast, in the bifactor model ESCS is not significantly correlated with the reading and science factors and its correlation with the math factors are small (not significant at ≤ 0.001) in both 2009 and 2010. The correlation between ESCS and the *g* factor are substantially larger: 0.37 and 0.36 in 2009 and 2010 respectively, indicating that ESCS is associated with general ability, but not with the specific learning domains.

TABLE 5

Table 5 shows that the associations between the Raven’s factor and the criterion variables are substantially smaller when it is isolated by the bifactor model rather than the four-dimensional model. In the multidimensional model, the Raven’s factor is positively correlated with enjoyment of reading in both years which is difficult to reconcile with the Raven’s test conceptualized as a non-verbal ability measure. Similarly, its negative relationship with being female is difficult to explain if Raven's is largely a measure of fluid intelligence. In contrast, these correlations are not found in the bifactor model. Furthermore, the positive correlation between Raven's and mathematics grades and its negative correlation with learning time in the humanities is consistent with its conceptualization as a non-verbal latent dimension.

Discussion

The aim of this study was to assess the degree to which students' PISA test scores reflect general cognitive ability rather than domain specific abilities. The analyses presented show that students' responses to the PISA items reflect mostly *g*. This study extends prior analyses based on German data suggesting that the multidimensional model employed to scale achievement data is outperformed by the bifactor model. Further analysis indicate that the four orthogonal factors (PISA reading, science, math and Raven's) are collectively responsible for around a quarter of the overall explained variance of the items but, each orthogonal factor only explains between 3% and 12% of the explained variance. The Raven's factor is far more reliable than the other specific ability factors. When the Raven's items are excluded, the general ability factor accounts for even more of the common variance in achievement (around 80% among the older students). The *Omega* and *PRMSE* indices confirm that a large part of the variation of subject specific achievement scores is driven by the general cognitive ability factor *g*. For the reading items, about 10% of the variation can be attributed to a specific, but unreliable, reading factor.

The orthogonal domain specific factors for reading, math and science estimated from the bifactor model are too unreliable to be considered substantially important. This may because the overarching goal of PISA is to assess general competencies across a range of real-life situations, although the items themselves are similar to items that assess schooling. It is possible that specific abilities would be more apparent from analyses of TIMSS data because TIMSS assesses curriculum-based skills and knowledge in mathematics and science. Saß, Kampa and Koller (2017) found that the multidimensional and bifactor models fitted TIMSS data equally well. An alternative model is that students' performance in all domains can be accounted for by a latent reading ability plus specific orthogonal factors for non-reading factors (so called bifactor-S-1 approach; see for instance Heinrich et. al. 2020). This model is suggested by the large correlations between reading and math and between reading and science in the PISA multidimensional models. Finally, students' performance in specific learning areas may be largely accounted for by general ability, whether the assessments are curriculum-based or not, and irrespective of the reading content of the specific test questions. The strong correlations between ability and achievement in meta-studies is consistent with this model and with the findings from this study, that variation student achievement can be attributed to a large extent to general ability, with specific abilities playing little role.

Criterion variables—grades, learning time, enjoyment of reading and gender—exhibit more theoretically consistent correlations with domain-specific factors isolated from the bifactor model in contrast to domain specific factors isolated from the standard multidimensional model. This is because the bifactor model generates purer specific ability factors and more substantively meaningful than the specific factor generated from the standard multidimensional model. Maximizing their domain-specific content and thus their reliability would require changes in item selection and item design.

Some conclusions drawn from analyses of PISA data may be unwarranted since students' scores from multidimensional models incorporate general ability. Policy makers have advocated changes to school curricula in mathematics to reduce socioeconomic inequalities in math demonstrated by analyses of PISA data (Schmidt, Burroughs, Zoido, & Houang, 2015; Schmidt, Zoido & Cogan, 2014). The finding that ESCS correlates more strongly with general ability than with reading or other specific abilities raises questions about its conceptualization and interpretations of its associations with student achievement. ESCS may incorporate parental cognitive abilities and early parental investments which, at least in part, accounts for the relationships between ESCS and the domain specific abilities generated from the standard multidimensional model.

This study suffers from limitations that can be addressed in future research. First, the PISA and Raven’s tests were not administered at the same time. Even though the correlations of PISA achievement domains across time is very high, an additional study with contemporaneous data may remove doubts that collecting data at multiple time points affected the results.

Second, the finding that LSAs are mainly about *g* are derived from analyses of only German and Polish data. Comparable analyses of LSA data from several countries would address whether the greater importance of *g* compared to domain specific factors is a general phenomenon. Parallel analyses that did not include the Raven’s items also found that most of the common variance was attributable to *g*. Therefore, comparisons of the importance of general and specific in PISA data from other countries does not require an accompanying cognitive ability test.

In this study there was a preponderance of reading items: 37 math items, 49 science items but 105 reading items were used. While the latent variable framework theoretically adjusts for varying numbers of items, further work using different proportions of items would confirm or refute the robustness of the latent structures identified here including the very weak math and science dimension.

Another limitation, which also applies to other ILSAs, is that in this study the multilevel structure of the data is not considered, i.e., that students are nested in classrooms and classrooms within schools. Ignoring the nested structure of our data should not affect analyses of dimensionality but might be relevant to the validation study of criterion variables. Although there are multilevel bi-factor models (Scherer & Gustafsson, 2015; Fujimoto, 2020) and general IRT models (Fox 2004), a multilevel approach for PISA items is beyond the scope of this study work and may be too computationally demanding.

This study could be expanded further to explore the dimensionality of the specific factors. For instance, the reading factor is the most variable, displaying non-trivial numbers of negative loadings. It is also the factor with the largest number of items because reading was the main domain in PISA 2009. Specifying reading as a single factor may not be sufficient, especially when the g-factor is controlled for, because of the greater variability in reading items than in mathematics or science items. This phenomenon may occur for whichever domain is the main domain in PISA; the greater number of test items produces greater heterogeneity. However, if the subdomains relate to items with common test stimuli, then they are likely to represent nuisance factors rather than reliable sub-domains. Exploring subdomains is a potential avenue for further research and refinement, bridging the work we developed at the item level with prior work that developed bifactor models using PISA subscales (Brunner, 2008; Baumert, et al., 2009).

This study found that students' responses to PISA test items reflect general ability rather than domain specific abilities. If this finding is replicated in other PISA data and with data from a range of achievement tests, it should prompt changes in test design and a shift in the interpretation of analyses of large-scale assessment studies. Specific ability factors isolated from bifactor models are very different from the corresponding factors derived from standard multidimensional models, substantially altering the relationships between predictor variables and specific abilities (for example the relation between socio-economic status and specific factors is close to zero). Our analyses suggest that a measurement model that considers the general ability factor fit PISA data substantially better than the multidimensional model that is routinely employed. Such model is congruent with theoretical work in both intelligence and achievement and is consistent with the strong correlations between ability and student achievement that have been documented in the literature. Therefore, the bifactor model should be incorporated into both item selection and analysis of LSAs. Such an approach would allow researchers to disentangle general cognitive abilities from subject specific abilities and enable the testing hypotheses of influences on specific abilities. General and specific subject abilities are both important for educational research and evidence-based educational policy.

Currently the measurement model in LSAs drives the design of assessment frameworks and data collection procedures. Field trial data are scaled, and items selected, based on how well they fit a specific ex-ante multidimensional scaling model that ignores the importance of the general factor identified in bifactor models. In order to maximize the informative value of the bifactor model for policy makers and educational researchers, that is to increase the reliability of domain-specific factors, a different item pool would need to be developed and different criteria for item selection. This would drive innovations in item design. Assessments of student performance in particular domains, item choice would be driven by the extent to which they load on specific ability factors rather than *g*. By contrast, assessments of general problem-solving abilities, item development and choice would be driven by the extent to which they load on *g*. Both these goals require moving from a multidimensional to bifactor modelling framework.

This study's findings, if replicated, would have major implications for the interpretation of analyses of achievement data. It would indicate that analysts need to be mindful of the contamination of the specific domains with general cognitive ability when considering relationships between test scores and socioeconomic, demographic, school, and teacher factors. Results indicate that differences in *g* are important in explaining differences in test scores isolated from standard models. Domain-specific scores generated from bifactor models, if reliable, could be more meaningfully linked to domain-specific covariates, for example, interest in science, enjoyment of reading, books in the home, teacher qualifications in math and science.

References

Agresti, A. (2010). Analysis of ordinal categorical data (Vol. 656). John Wiley & Sons.

Allik J., Asendorpf J.B., Bosker R.J., Brunner M., Martin R., Ceci S.J., Williams W.M., De Fruyt F., Demetriou A., Flynn J.R., Helfrich H., Hunt E., Irwing P., Johanson W., Lynn R, Meisenberg G., … , Wicherts J.M., & Wilhelm O. (2007), Discussion on 'The *g*-Factor of International Cognitive Ability Comparisons: The Homogeneity of Results in PISA, TIMSS, PIRLS and IQ-Tests Across Nations' by Heiner Rindermann. *European Journal of Personality., 21,* 707-765.

Avvisati, F. (2020). The measure of socio-economic status in PISA: a review and some suggested improvements. *Large Scale Assessments in Education, 8.*

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale Student Assessment Studies Measure the results of Processes of Knowledge Acquisition: Evidence in Support of the Distinction between Intelligence and Student Achievement. *Educational Research Review*, *4*(3), 165-176.

Bates, Timothy C., et al. 2019. "Social Competence in Parents Increases Children's Educational Attainment: Replicable Genetically- Mediated Effects of Parenting Revealed by Non-Transmitted DNA." Twin Research and Human Genetics 22 (1): 1-3.

Bates, Timothy C., et al. 2018. "The Nature of Nurture: Using a Virtual-Parent Design to Test Parenting Effects on Children's Educational Attainment in Genotyped Families." Twin Research and Human Genetics 21 (2): 73-83.

Buchmann, Claudia. 2002. "Measuring family background in international studies of education: Conceptual issues and methodological challenges." In Methodological advances in cross-national surveys of educational achievement, edited by A.C. Porter and Adam Gamoran, 150–197. Washington, DC: National Academy Press.

Beaujean, A. A., Parkin, J., & Parker, S. (2014). Comparing Cattell–Horn–Carroll factor models: Differences between bifactor and higher order factor models in predicting language achievement. *Psychological Assessment*, *26*(3), 789.

Bentler, P. M. (2009). Alpha, Dimension-Free, and Model-Based Internal Consistency Reliability. *Psychometrika*, *74*,(1),137–43.

Berge, Jos M. F. ten, & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*(4), 613–25.

Bol, T., & Van de Werfhorst, H. (2013). The measurement of tracking, vocational orientation, and standardization of educational systems: A comparative approach. *AIAS: GINI Discussion Paper 81*.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental Measurement in the Social Sciences. Mahwah,* NJ: Lawrence Erlbaum Associates.

Borgonovi, F., & Pokropek, A. (2019). Seeing is believing: Task-exposure specificity and the development of mathematics self-efficacy evaluations. *Journal of Educational Psychology*, *111*(2), 268.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers*, No. 71, OECD Publishing, Paris.

Brunner, M. (2008). No *g* in education?. *Learning and Individual Differences, 18*(2), 152-165.

Carroll, J. B. (1993). *Human Cognitive abilities: A Survey of Factor-Analytic Studies.* New York: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1-22.

Coburn, C. E., Hill, H. C., & Spillane, J. P. (2016). Alignment and accountability in policy design and implementation: The Common Core State Standards and implementation research. *Educational Researcher*, *45*(4), 243-251.

Cromley, J. G. (2009). Reading Achievement and Science Proficiency: International Comparisons from the Programme on International Student Assessment. *Reading Psychology*, *30*(2), 89-118.

Cucina, J., & Byle, K. (2017). The Bifactor Model Fits Better Than the Higher-Order Model in More Than 90% of Comparisons for Mental Abilities Test Batteries. *Journal of Intelligence*, *5*(3), 27-27.

Deng, Z., & Gopinathan, S. (2016). PISA and high-performing education systems: Explaining Singapore’s education success. *Comparative Education*, *52*(4), 449-472.

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, 76(5), 741-770.

Egelund, N. (2008). The value of international comparative studies of achievement–a Danish perspective. *Assessment in Education: Principles, Policy and Practice, 15*(3), 245–251.

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor Models for Predicting Criteria by General and Specific Factors: Problems of Nonidentifiability and Alternative Solutions. *Journal of Intelligence, 6*(3), 42.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127.

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education, 32*(5), 619–634.

Floyd, R. G., Reynolds, M. R., Farmer, R. L. & Kranzler, J. H. (2013). Are the general factors from different child and adolescent intelligence tests the same? Results from a five-sample, six-test analysis, *School Psychology Review, 42*(4), 383—401.

Fox, J. P. (2004). Applications of multilevel IRT modeling. School Effectiveness and School Improvement, 15(3-4), 261-280.

Fujimoto, K. A. (2020). A More Flexible Bayesian Multilevel Bifactor Item Response Theory Model. Journal of Educational Measurement, 57(2), 255-285.

Gignac, G E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence, 62,*138–47. doi: 10.1016/j.intell.2017.04.001.

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. Multivariate Behavioral Research, 48(5), 639-662.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for *g* factor theory and the brief measurement of g. *Intelligence, 52,* 71-79.

Gottfredson, L. S. (1997). Mainstream Science on Intelligence: An Editorial with 52 signatories, History, and Bibliography. *Intelligence*, *24*(1), 13-23.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60(1)*, 549-576.

Grek, S. (2009). Governing by numbers: the PISA 'effect' in Europe. *Journal of Education Policy*, *24*(1), 23-37.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407-434.

Gustafsson, J. E., & Undheim, J. O. (1996). Individual differences in cognitive functions. Pp. 186-242. In (Eds.) Berliner, D.C., & Calfee, R. C. *Handbook of Educational Psychology*. New York: Macmillan.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204-229.

Haberman, S., Sinharay, S., & Puhan, G. (2008). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*, 79-95.

Halpern, D. F. (2012). Sex Differences in cognitive abilities. 4th Edition. New York, Psychology Press.

Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2020). Giving G a meaning: An application of the bifactor-(S-1) approach to realize a more symptom-oriented modeling of the Beck depression inventory–II. *Assessment*, *27*(7), 1429-1447.

Hopfenbeck, T.N., and Kjærnsli, M. (2016). Students' test motivation in PISA: the case of Norway, *The Curriculum Journal*, *27*(3), 406-422.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Hunt, E. (2010). *Human intelligence*. Cambridge University Press.

Jakubowski, M., & Pokropek, A. (2015). Reading achievement progress across countries. *International Journal of Educational Development*, *45*, 77-88.

Jaworowska, A., Szustrowa, T., & Raven, J. C. (2000). *Test Matryc Ravena w wersji Standard TMS: formy: Klasyczna, Równoległa, Plus: polskie standaryzacje*. Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Jensen A.R. and Li-Jen Weng (1994) What is a good *g*? *Intelligence*, *18*(3), 231-258. https://doi.org/10.1016/0160-2896(94)90029-9.

Johnson, W., Bouchard, T. J., Jr., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one *g*: consistent results from three test batteries. *Intelligence, 32*(1), 95-107.

Johnson, W., Nijenhuis, J. T., & Bouchard, T. J., Jr. (2008). Still just 1 *g*: Consistent results from five test batteries. *Intelligence*, *36*(1), 81-95.

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive *g* and academic achievement *g* one and the same *g*? An exploration on the Woodcock–Johnson and Kaufman tests. *Intelligence*, *40*, 123–138.

Keller, L., Preckel, F., & Brunner, M. (2020). Nonlinear relations between achievement and academic self-concepts in elementary and secondary school: An integrative data analysis across 13 countries. *Journal of Educational Psychology.* Advance online publication.

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence*, *36*, 153–160.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904)"'General intelligence,'objectively determined and measured". *Journal of Personality and Social Psychology, 86*(1), 96.

Lynn, R., & Vanhanen, T. (2012). *Intelligence: A unifying construct for the social sciences*. Ulster Institute for Social Research.

Lynn, R., Harvey, J., & Nyborg, H. (2009). Average intelligence predicts atheism rates across 137 nations. *Intelligence, 37*(1), 11-15.

Martin, M. O., Mullis, I. V., & Hooper, M. (2017). *Methods and Procedures in PIRLS 2016*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College International Association for the Evaluation of Educational Achievement.

McDonald, R. P. (1999). Test theory: A unified approach. Mahwah, NJ: Erlbaum.

Marks, G. N. (2016) Explaining the substantial inter-domain and over-time correlations in student achievement: The importance of stable student attributes, *Educational Research and Evaluation, 22*(1–2), 45–64.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*(2), 19four-199.

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User’s Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and Unknowns. *American Psychologist*, *51*(2), 77-101. doi:10.1037/0003-066X.51.2.77

OECD. (2007). Science Competencies for Tomorrow's World. (Vol. 1). Paris: Organisation for Economic Co-operation and Development.

OECD. (2009). PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science. Paris: Organisation for Economic Co-operation and Development.

OECD. (2012). PISA 2009 Technical Report. Paris: OECD Publishing.

OECD. (2013). PISA 2012 Results: Excellence through Equity Giving Every Student the Chance to Succeed (Vol. II). Paris: OECD Publishing.

OECD. (2014). PISA 2012 Technical Report. Paris: OECD Publishing.

OECD. (2016). PISA 2015 Results (Volume I). Paris: OECD Publishing.

Raven, J. (2003). Raven Progressive Matrices. In R. S. McCallum (Ed.), Handbook of Nonverbal Assessment (pp. 223-237). Boston, MA: Springer US.

Raykov, T. (1997). Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement*, *21(2)*, 173–184.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment,* 95, 129–140.

Reise, S. P., Bonifay, W., & Haviland, M. G. (2018). Bifactor Modelling and the Evaluation of Scale Scores. pp. 675–707. in (Eds.) Irwing, P. Booth, T., & Hughes, D.J. *The Wiley* *handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. Chichester: Wiley Blackwell.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*(1), 5-26.

Rindermann, H. (2007). The *g*-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality: Published for the European Association of Personality Psychology*, *21*(5), 667-706.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. *Journal of Personality Assessment*, *98*(3), 223-237. doi:10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137-150.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal, 4*(3), 227–241.

Saß, S., Kampa, N., & Köller, O. (2017). The interplay of *g* and mathematical abilities in large-scale assessments across grades. *Intelligence*, *63*, 33-44.

Scherer, R., & Gustafsson, J. E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: An application of multilevel bifactor structural equation modeling. *Frontiers in psychology*, 6, 1550.

Schleicher, A. (2007) Can Competencies Assessed by PISA Be Considered the Fundamental School Knowledge 15-Year-Olds Should Possess?. *Journal of Educational Change, 8*(4), 349-357.

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher*, *44*(7), 371-386.

Schmidt, W. H., Zoido, P., & Cogan, L. (2014). Schooling matters: Opportunity to learn in PISA 2012. *OECD Education Working Papers*, No. 95, OECD Publishing, Paris.

Sellar, S., and Lingard, B. (2013). The OECD and the expansion of PISA: new global modes of governance in education. *British Educational Research Journal, 40*(6), 917-936.

Spearman, C. (1904). “General intelligence,” objectively determined and measured. *American Journal of Psychology, 15,* 201–292.

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. PLOS ONE, 8(3), e57988. Retrieved from http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0057988

Takayama, K. (2008). The politics of international league tables: PISA in Japan’s achievement crisis debate. *Comparative Education*, *44*(4), 387–407.

Van Hek, M., Buchmann, C., & Kraaykamp, G. (2019). Educational systems and gender differences in reading: A comparative multilevel analysis. *European Sociological Review*, *35*(2), 169–186. https://doi.org/10.1093/esr/jcy054

Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership*, *41*(8), 19–27.

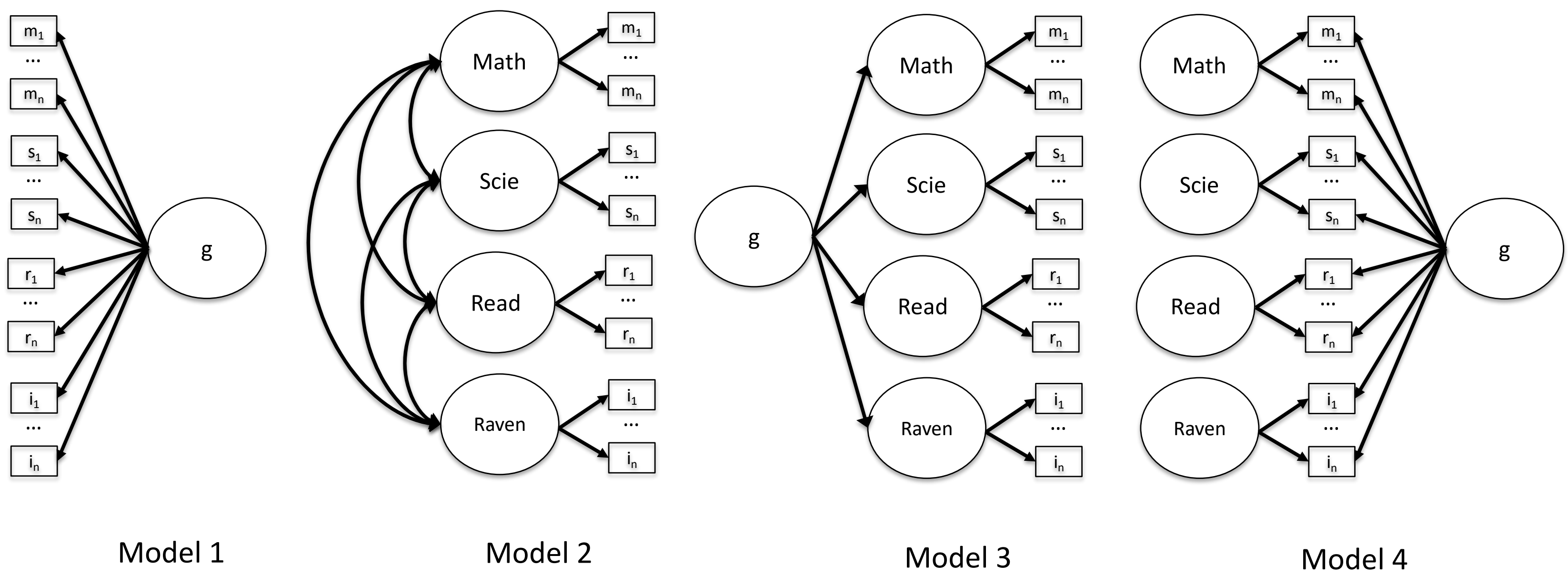
Warne, R. T., & Burningham, C. (2019). Spearman’s *g* found in 31 non-western nations: strong evidence that *g* is a universal phenomenon. *Psychological Bulletin*, *145*, 237–72.

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113-128.

Zaboski II, B. A., Kranzler, J. H., & Gage, N. A. (2018). Meta-analysis of the relationship between academic achievement and broad abilities of the Cattell-Horn-Carroll theory. *Journal of school psychology*, *71*, 42-56.

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change*, *21*, 245-266. https://doi.org/10.1007/s10833-019-09367-x

Figure 1. Potential relations between academic abilities and intelligence measure by Raven's test



Model 1 - One-dimensional IRT model; Model 2 - Four-dimensional IRT model; Model 3 - Higher-Order IRT; Model 4 – Bifactor IRT model; *m, s, r, i* indicate individual math, science, reading and Raven's items (responses to items are treated as categorical variables).

Figure 2. Item loadings form Bifactor IRT 2009 model.

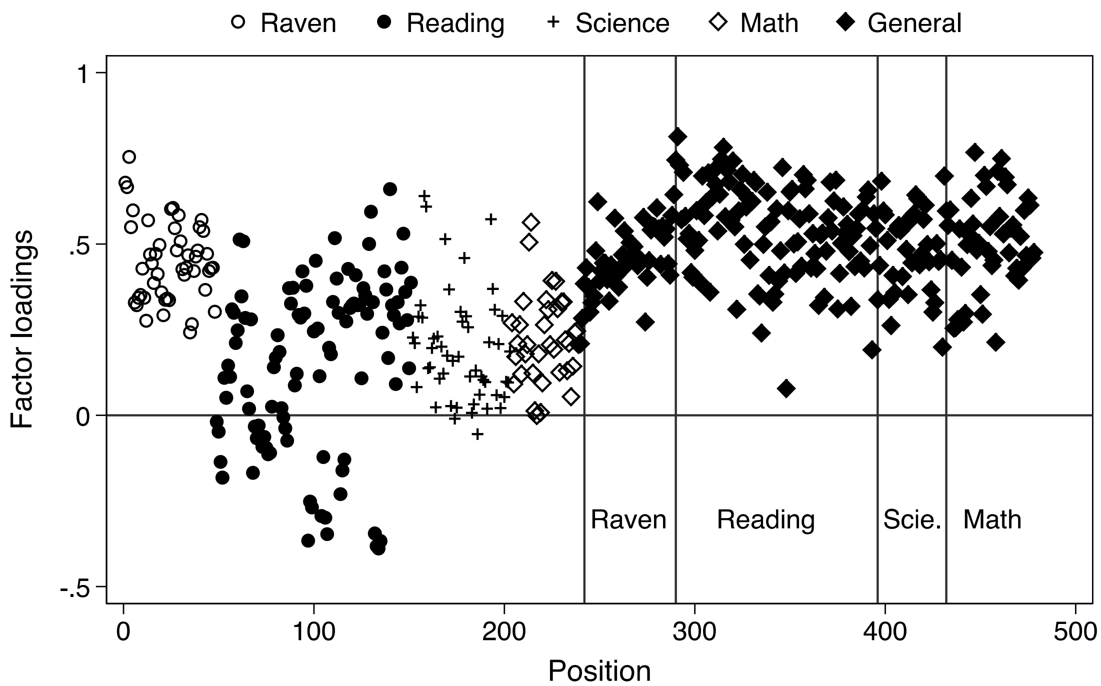


Figure 3. Item loadings form Bifactor IRT 2010 model.

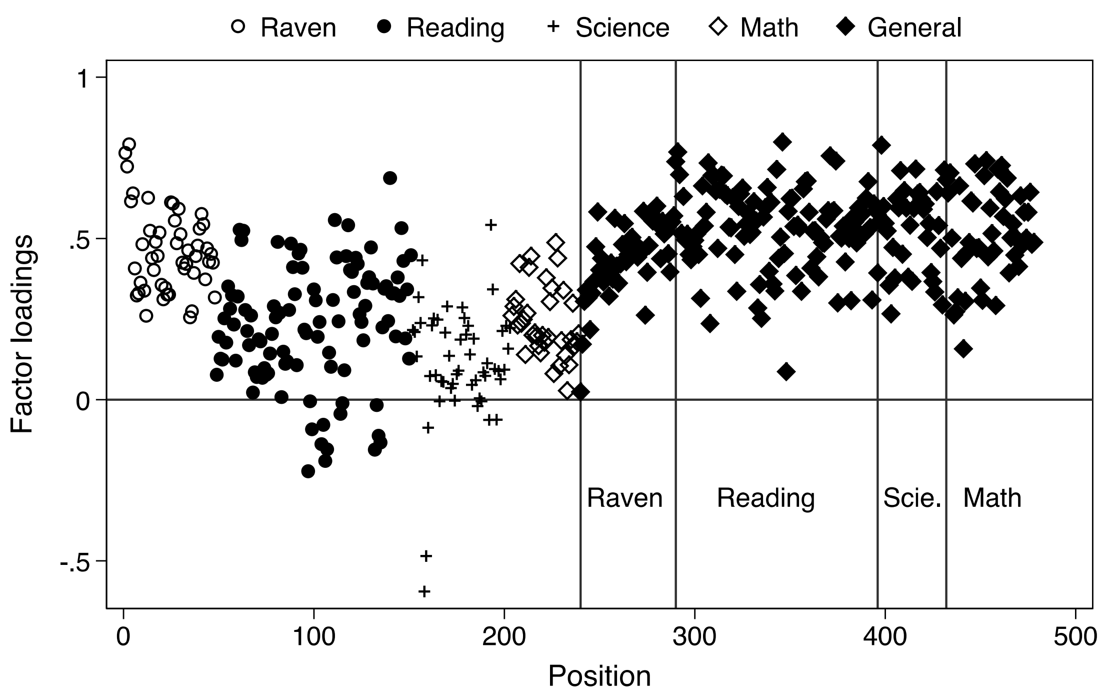


Table 1

Pearson correlations between latent traits from Four-dimensional model.

Ability	Raven	PISA Reading	PISA Science
		2009	
Raven	1		
PISA Reading	0.70 (0.01)	1	
PISA Science	0.69 (0.02)	0.86 (0.01)	1
PISA Mathematics	0.80 (0.01)	0.82 (0.01)	0.89 (0.01)
		2010	
Raven	1		

PISA Reading	0.69 (0.01)	1	
PISA Science	0.70 (0.87)	0.87 (0.01)	1
PISA Mathematics	0.81 (0.01)	0.82 (0.01)	0.90 (0.01)

Note. Standard errors in brackets.

Table 2

Model fit for models based on 2009 and 2010 measurement

Year	Variables	Model 1	Model 2	Model 3	Model 4
	Number of free parameters	478	484	482	717
	Log-likelihood	-213981	-209984	-210038	-189529
2009	Akaike Information Criterion (AIC)	428918	420936	421039	380493
	Bayesian Information Criterion (BIC)	432050	424107	424197	385094
	Sample-Size Adjusted BIC	430531	422569	422665	382815
	Number of free parameters	478	484	482	717

	Log-likelihood	-184727	-180872	-180927	-158678
2010	Akaike Information Criterion (AIC)	370410	362712	362817	318789
	Bayesian Information Criterion (BIC)	373462	365803	365896	323276
	Sample-Size Adjusted BIC	371944	364265	364364	320998

Table 3

Descriptive statistics for item loadings.

	2009				2010			
Factor	mean	std	min	max	mean	sd	min	max
General	0.51	0.13	0.08	0.81	0.51	0.13	0.02	0.80
Math	0.22	0.13	-0.001	0.56	0.24	0.11	0.03	0.49
Raven	0.45	0.12	0.24	0.75	0.46	0.13	0.25	0.79
Reading	0.15	0.25	-0.39	0.66	0.23	0.19	-0.22	0.69
Science	0.19	0.16	-0.05	0.64	0.12	0.18	-0.59	0.54

Table 4

General factor strength indices

Year	Dimension	EVC	Omega	Omega (H/HS)	Omega ratio	PRMSE subscale	PRMSE total	PRMSE ratio
2009	General	0.73	0.99	0.94	0.94	-	-	-
	Math	0.04	0.95	0.14	0.15	0.83	0.81	1.03
	Raven	0.12	0.97	0.49	0.50	0.90	0.61	1.48
	Reading	0.03	0.98	0.07	0.07	0.90	0.83	1.09
	Science	0.10	0.95	0.14	0.15	0.84	0.76	1.11
2010	General	0.73	0.99	0.92	0.93	-	-	-
	Math	0.03	0.96	0.05	0.05	0.85	0.86	0.98
	Raven	0.12	0.97	0.52	0.54	0.90	0.58	1.55

Reading	0.10	0.98	0.16	0.16	0.91	0.83	1.10
Science	0.03	0.95	0.16	0.17	0.85	0.76	1.11

Note: ECV - Explained Common Variance; PRMSE proportional reduction in mean squared error based on subscale scores. See online appendix for formulae.

Table 5

Correlations between latent variables and validation variables in 2009 and 2010 measurement models

Year	Variables	Reading		Science		Math		Raven		General
		Multi-d	Bifactor	Multi-d	Bifactor	Multi-d	Bifactor	Multi-d	Bifactor	Bifactor
2009	Language grades	0.36***	0.05*	0.30***	0.00	0.26***	-0.06	0.23***	-0.01	0.38***
	Mathematics grades	0.36***	0.01	0.33***	-0.07*	0.46***	0.14***	0.39***	0.12***	0.42***
	Biology grades	0.32***	0.07*	0.32***	0.02	0.31***	0.08*	0.24***	0.01	0.35***
	Enjoyment of reading	0.33***	0.13**	0.24***	-0.02	0.17***	-0.06*	0.16***	-0.03	0.34***
	ESCS	0.25***	0.03	0.26***	0.02	0.29***	0.07*	0.22***	0.01	0.37***
	Learning time humanities	0.25***	0.13***	0.19***	-0.03	0.18***	0.00	0.07	-0.02	0.49***
	Learning time math	0.24***	0.08***	0.20***	0.00	0.25***	0.03	0.27***	0.11***	0.45***
	Learning time science	0.33***	0.08***	0.30***	0.01	0.33***	0.03	0.26***	0.03	0.53***
	Female	0.14***	0.13***	-0.09*	-0.19***	-0.14***	-0.19***	0.00	-0.02	0.09**
2010	Language grades	0.39***	0.15***	0.29***	-0.04	0.32***	-0.01	0.23***	-0.02	0.38***
	Mathematics grades	0.36***	0.08**	0.34***	0.01	0.49***	0.27***	0.39***	0.14***	0.40***
	Biology grades	0.30***	0.09**	0.27***	-0.03	0.32***	0.07	0.24***	0.02	0.34***
	Enjoyment of reading	0.34***	0.16***	0.26***	0.02	0.17***	-0.20***	0.15***	-0.03*	0.36***
	ESCS	0.22***	0.01	0.25***	0.01	0.29***	0.09**	0.24***	0.02**	0.36***
	Learning time humanities	0.32***	0.20***	0.25***	-0.10**	0.17**	-0.09*	0.05	-0.04	0.55***
	Learning time math	0.16***	0.06**	0.17***	-0.05	0.26***	0.09	0.25***	0.08	0.40***
	Learning time science	0.35***	0.16***	0.30***	-0.04	0.33***	0.00	0.29***	0.01	0.53***
	Female	0.14***	0.27***	-0.09*	-0.22***	-0.14***	-0.35***	0.00	-0.01	0.10**

Note: ***p ≤ 0.001; **p<0.01; *p≤0.05. four-d refers to the four-dimensional model.