General or specific abilities? Evidence from 33 countries participating in the PISA assessments

Published in: Intelligence 92, May-June 2022 https://doi.org/10.1016/j.intell.2022.101653

Artur Pokropek [a,*], Gary N. Marks [b], Francesca Borgonovi [c], Piotr Koc [d], Samuel Greiff [e]

[a] *Educational Research Institute (IBE), Górczewska 8, 01-180 Warsaw, Poland*

[b] *Department of Sociology, Social and Political Sciences, University of Melbourne, Melbourne, Vic, Australia*

[c] *UCL Social Research Institute, University College London, 55-59 Gordon Square, London WC1H 0NU, United Kingdom*

[d] *Institute of Philosophy and Sociology of the Polish Academy of Sciences, 72 Nowy Świat Street, 00-330 Warsaw, Poland*

[e] *Cognitive Science and Assessment, ECCS unit, University of Luxembourg, 11, Porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg*

Abstract

Psychometricians working on International Large Scale Assessments (ILSAs) typically specify latent ability factors with distinct and correlated constructs for test domains, such as reading, mathematics and science. A construct for general ability is not specified. However, several country-specific studies conclude that ILSAs largely reflect general ability. We extend such studies and examine the dimensionality of the 2018 PISA assessment in 33 OECD countries examining three models: three-dimensional IRT model, the bifactor IRT model and the bifactor (S-1) IRT model. A four-tiered approach was adopted. First, models were compared using an information criterion (AIC). Second, the correlations from the multidimensional model were estimated to assess in which countries the three dimensions are sufficient discriminant validity. Third, a variety of bifactor indices were utilized to establish the explanatory power and reliabilities of the latent dimensions generated by the three models. Finally, the statistical relationships between the latent factors derived from the three models and educationally relevant covariates were estimated. The bifactor model fits the data better than standard multi- dimensional model or S-1 model in every country investigated. The correlations in the correlated factor model are above 0.8 in all 33 countries. The symmetrical bifactor general ability model shows that 80%, or more, of the common variance in student responses to the PISA instruments is accounted for by a general ability factor. On average, 27% of variance in the mathematics items is independent of the general factor and can be attributed to a specific mathematics ability factor. The respective estimates for reading are 12% and science is 17%. Re- lationships for selected covariates with the PISA domains follow the same pattern as general ability in the bifactor model.

1. Introduction

There is an ongoing debate whether prominent achievement and aptitude instruments measure knowledge and skills for specific subject domains or general ability. Some studies conclude that achievement tests mostly measure general ability rather than domain-specific abilities that they purport to measure, such as the Scholastic Assessment Test (SAT), American College Testing (ACT), the OECD's Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS) studies (Frey & Detterman, 2004; Koenig, Frey, & Detterman, 2008; see also Luo, Thompson, & Detterman, 2003). In this paper we focus on PISA, the most prominent International Large-Scale Assessment of Student Achievement (ILSA). PISA has two main goals. The first is to assess how well students can apply their knowledge and skills to solve problems in real-life situations. PISA assessments are informed but are not constrained, by the national

curricula of participating countries (OECD, 2019a, p. 9). The second goal of PISA is the accountability of educational systems: the extent that students meet expected standards, the performance of disadvantaged demographic and socioeconomic groups, and school and teacher differences (Caro, Lenkeit, & Kyriakides, 2016; Jakubowski & Pokropek, 2015).

PISA is important to educational policy. For example, the European Union's strategic framework for co-operation in education and training includes indicators from PISA (European Union, 2009). Similarly, the framework for monitoring progress towards the United Nations' Sustainable Development Goals includes PISA, together with indicators from other international assessments (UNESCO, 2019). PISA has been used to justify educational reforms in several countries (Ertl, 2006; Grek, 2009; Breakspear, 2012; Takayama, 2008; Dobbins & Martens, 2012) PISA assesses three core domains: mathematics, reading, and science. Occasionally, additional domains are included, for example problem solving, financial literacy and creative thinking. The results for each dimension are reported separately consistent with the study's fundamental assumptions that PISA measures the subject domains reli- ably and that they are largely independent.

This paper examines to what extent the instruments in the PISA 2018 study met these assumptions by examining the role of general ability in shaping individual variation in response patterns in the PISA test. The general ability factor is generally referred to as 'general academic ability' by education specialists and 'intelligence' by psychologists. The primary purpose of the paper is to investigate to what extent the PISA 2018 assessment measured general ability vis-à-vis domain-specific knowledge and skills, analysing data from 33 OECD countries. This research question is important to educational research, policy and the design of assessment instruments, not just for PISA but for student achievement studies in general.

ILSAs typically conceptualise domain-specific abilities in multidimensional models as single latent variables, such as reading, mathematics and science (as in PISA). Each assessment domain corresponds to a single latent factor which supposedly fully represents capability in the specific domain (OECD, 2019b). The latent constructs are specified as correlated. In this approach, general cognitive ability is not modelled and as an important omitted variable is likely to be confounded with domain-specific factors.

In contrast, a bifactor model for ILSAs specifies that variation in student responses is accounted for by general ability, together with in- dependent domain-specific abilities. In such models, each item loads on general ability, and domain specific items (e.g., math) also load on their respective specific ability factor. Studies on data from German and Polish students find that the bifactor model fits student achievement data better than the multidimensional model and yields more plausible relations with socio-economic and demographic characteristics, and educational inputs (Baumert, Lüdtke, Trautwein, & Brunner, 2009; Saß, Kampa, & Köller, 2017; Pokropek, Marks, & Borgonovi, 2021). In those studies, domain-specific abilities exhibited some small additional explanatory power, suggesting that ILSAs mainly measure general ability together with less reliably measured scholastic abilities. How- ever, bifactor models possess a strong tendency to fit any possible data due to their inherent ability to capture random noise in the data (Bonifay & Cai, 2017).

Alternatively, student responses could be governed largely by their reading ability rather than their general ability. In the PISA mathematics and science assessments, students are presented with textual material which they must comprehend and interpret to be able to answer the associated questions correctly (OECD, 2019a). Given the importance of reading for PISA, an appropriate model may be the S-1 bifactor model in which reading is assumed to be the general ability that influences the responses to PISA items in all domains, together with independent domain-specific abilities in mathematics and science (Eid, Krumm, Koch, & Schulze, 2018; Heinrich, Zagorscak, Eid, & Knaevelsrud, 2020). Previous studies of this issue have been limited to two countries: Germany and Poland. Primarily two types of models were compared: the multidimensional and symmetrical bifactor models which includes a general ability factor. The present study substantially increases the scope of previous studies by investigating the dimensionality of students' responses to PISA in 33 OECD countries (listed in the tables) and adds the non-symmetrical bifactor S-1 bifactor reading model, which replaces the general ability factor with a general reading factor. The first group of analyses compare the fit of three measurement models: the standard multidimensional model typically used in national and international assessments, the symmetrical bifactor model, and the bifactor reading model. The second group of analyses evaluates the extent that relationships of the latent factors isolated with several

demographic, socioeconomic and educational variables are consistent with theoretical expectations.

This study analyses the *nomological* network of PISA, that is, representations of the concepts (constructs), their observable manifestations, and their interrelationships. Like any statistical model, the models used in this study are used simplify the relationships between variables, ultimately, the individual test items in the PISA study. Such models cannot fully reflect the complex relationships between students' abilities and test items. They are used instrumentally, following the well-known adage *that all models are wrong but some of them are useful*(Box, 1976). Investigating this set of models, addresses the central questions posed in this paper: does PISA largely measure general ability or the purported domain-specific abilities of reading, mathematics and science, and are their relationships with exogenous variables consistent with theoretical expectations?

1.1. Evidence that cognitive ability is relevant to student achievement

There are five research literatures that suggest that much of the variation in student responses to ILSAs, such as PISA can be accounted for by general cognitive:

(1) Conceptualizations of literacy in PISA and general ability.
(2) Correlations between student achievement and intelligence tests.
(3) Inter-domain correlations in student achievement.
(4) Behavioural Genetics
(5) Psychometric modelling.

1.1.1. Conceptualization of 'literacy' in PISA and general ability

Formally, PISA is designed to measure students' capacity to apply knowledge and skills in key subject areas and to analyse, reason and communicate effectively as they pose, solve and interpret problems in a variety of situations that can take place in a mathematics, reading, or science context (OECD, 2007, p.16). This definition is almost identical to prominent definitions of intelligence, i.e., the 'ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought' (Neisser et al., 1996, p. 77).Armor (2003, p. 19) noted that both achievement tests and intelligence tests include similar subsets of items, for example assessing vocabulary, reading comprehension, mathematical concepts, numerical skills. He suggests that the substantial overlap between IQ and achievement tests indicates they are measuring something in common: general reasoning skills.

General ability is likely to be more important in PISA, compared to other system-wide international student assessments, such as the Trends in Mathematics and Science Study (TIMSS) and the Progress in Inter- national Reading Literacy Study (PIRLS). These assessments are based on curricular content, whereas PISA aims to assess the extent to which education systems as well as the social and economic context children experience up to age 15 equip young people with general life skills rather than specific knowledge and skills taught at school (Egelund, 2008; Schleicher, 2007). Rindermann and Baumeister (2015) used 67 expert raters to assess the abilities required to correctly answer PISA and TIMSS items and found that general intelligence and general knowledge were rated as important components, more so for PISA than TIMSS.

1.1.2. Correlations between student achievement and intelligence tests

On country-level relations, Rindermann (2007) concluded that assessments of student achievement and intelligence tests both measure general cognitive ability highlighting the role of social and economic context in shaping general ability. Rindermann found strong correlations between 20 country-level student achievement measures and country-level intelligence measured by Lynn and Vanhanen (2012). However, correlations at the country-level do not necessarily individual- level correlations and vice versa (Robinson, 1950).

The individual-level correlations between standardised achievement tests and intelligence range from about 0.50 to over 0.80, varying by subject domain, students' age or grade-level, test reliability and educational and social context. In a meta study of close to 3000 empirical studies of school learning in the Unites States, Walberg (1984) computed an average correlation of 0.71 between various IQ measures and academic achievement. For the US, Duckworth, Quinn, and Tsukayama (2012) reported correlations between 0.70 and 0.80 for IQ measured in grade 4, and grade 5 and 9 achievement tests. For Australia, the correlations of early cognitive ability with numeracy and reading were around

0.60 for Year 3 students (Marks, 2016). For New Zealand, the correlation between IQ measured at ages 8 and 9 with academic per- formance at age 13 was 0.83 (Fergusson, Horwood, & Boden, 2008). For Ireland, the correlations between cognitive ability measured at age 13 and numeracy and vocabulary at age 17 were around 0.60 (O'Connell & Marks, 2021). For Germany, Baumert, Nagy and Lehmann (2012, Table S1) reported correlations around 0.50 for IQ with both reading and mathematics. For the Netherlands, Bartels, Rietveld, Van Baal, and Boomsma (2002) reported correlations of 0.41, 0.50, 0.60, and 0.63 between IQ assessed at age 5, 7, 10, and 12 respectively performance in the CITO tests (which influences school track placement) taken at 12 years of age. For Sweden, the correlation between cognitive ability measured at age 13 and grades is around 0.60 among cohorts born from 1967 to 1982 (Erikson & Rudolphi, 2010).

Structural equation models (SEM), which account for measurement error, almost invariably produce higher correlations than correlations between manifest variables. In a study of 178,599 pupils attending En- glish state schools, the correlation between latent factors derived from a cognitive ability test and attainment scores on national Key Stage 2 tests in English, mathematics and science of 11-year-olds was 0.83 (Calvin, Fernandes, Smith, Visscher, & Deary, 2010). In a study of over 80,000 16-year-old students, Deary, Strand, Smith, and Fernandes (2007) calculated a correlation of 0.81 between a latent intelligence trait measured at 11 years of age with a latent trait of subject performance in the General Certificate of School Education taken around age 16. Zaboski, Kranzler, and Gage's (2018) metastudy estimated correlations of between 0.70 and 0.80 between *g* and basic reading skills, reading comprehension and basic math. General intelligence *g* extracted from the Armed Services Vocational Aptitude Battery correlated at around
0.80 with the Scholastic Assessment Test and the American College Readiness Assessment (Baker et al., 2004; Frey, 2019; Koenig et al., 2008).

According to Hopfenbeck et al. (2018, pg. 342) 54 articles have investigated the relationship between intelligence and PISA. They offer an overall correlation of 0.88 implying that, on average, about 80% of the variation is the PISA can be attributed to cognitive ability. Other estimates have been lower. For Poland, the latent correlations between Raven's progressive matrices, a measure of fluid intelligence, and PISA test scores in Poland were around 0.73 (Pokropek et al., 2021).

1.1.3. High inter-domain correlations in student achievement

The origin of concept general ability is attributed to Spearman's (1904) identification a general intelligence (or general ability) factor from students' performance in different subject areas using factor anal- ysis. The general intelligence factor, now known as *g*, is a latent variable that explained the correlations, among a diverse set of school subjects. All cognitive tasks measure *g* to some degree (Warne, 2020, p. 32) and, ideally, the goal of education systems is to develop students' general ability to solve cognitive tasks in real world settings.

For student performance in achievement tests, reading scores tend to be highly correlated with student performance in other domains. In a French-Canadian study of school readiness, the correlation between grade 2 reading and math was 0.75 (Pagani, Fitzpatrick, Archambault, & Janosz, 2010). In 2011, over 4000 grade 4 Italian students were tested in reading in PIRLS, and in math and science in TIMSS. Reading scores were highly correlated with math (0.76) and science (0.85) score com- parable with the correlation (0.81) between science and math scores in TIMSS (Grilli, Pennoni, Rampichini, & Romeo, 2016). For Australia, Marks (2021) reported interdomain correlations of numeracy with 4 English literacy domains between 0.5 and 0.8. The most plausible explanation for the common variance in these studies is general cognitive ability.

For PISA, correlations between the PISA factors derived from the multidimensional model are very high, typically between 0.80 and 0.90 (Bond & Fox, 2001; Cromley, 2009). According to the OECD's PISA (2019a, Table 12.14) technical report on the 2018 PISA assessments the inter-domain correlations were also very high. Across countries with computer-based assessments, the average correlation between mathematics and reading was 0.80, 0.82 between mathematics and science and 0.86 between reading and science. According to Brown (2015), highly correlated dimensions (i.e., above 0.80) is indicative of poor discriminant validity. The separate dimensions cease to be meaningful measures of separate domains, but imperfect measures of the same general latent construct. Again, general ability is the most obvious explanation for the high correlations between students' scores in different achievement domains derived from the standard multidimensional model. The very

high inter-domain correlations means that analyses of the separate domains will produce very similar results because of the highly correlated latent dimensions, with the notable exception of gender differences.

1.1.4. Behavioural genetics

Within homogeneous populations, the literature indicates that the genetic component of student achievement is comparable to, or greater than, that for cognitive ability (Kovas et al., 2013). The heritabilities, that is the proportions of variation in traits among homogeneous populations that can be attributed to genes, are generally between 0.5 and 0.8, averaging about 0.7, with much lower estimates for the shared environment (see Plomin, DeFries, Knopik, & Neiderhiser, 2013, pp. 222–228; Pokropek & Sikora, 2015; Grasby, Coventry, Byrne, Olson, & Medland, 2016). The shared environment encompasses factors such as family background and schools. A meta-analysis of 61 twin studies from 11 cohorts of primary school children reported heritabilites ranging from 0.4 to 0.7, whereas the contributions of the shared environment were mostly around 0.10 (de Zeeuw, de Geus, & Boomsma, 2015). Asbury and Plomin (2014, pp. 39, 45) cited heritability estimates of around 0.6 for reading, and between 0.6 and 0.7 for math. It is critical however to consider that the low variation attributed to environmental factors pertains estimates of highly homogeneous populations, as in the example of twin studies, and therefore should not be extrapolated to indicate the role of environmental factors when comparing heterogeneous populations, such as students in different countries (Paige Harden, 2021).

Empirical studies estimated sizable genetic correlations between achievement domains and cognitive ability (Hart, Petrill, Thompson, & Plomin, 2009; Petrill, 2016; Wainwright, Wright, Luciano, Geffen, & Martin, 2005). The Generalist Genes Hypothesis posits that the same set of genes largely shape a wide range of cognitive and learning abilities. Kovas, Harlaar, Petrill, and Plomin (2005) estimated an average genetic correlation between achievement domains of 0.79 and genetic correlations within homogeneous populations between 0.47 and 0.76 for *g* and the achievement domains.

1.1.5. Psychometric modelling

The psychometric structures of student responses to assessments suggest general ability factors. Reise, Bonifay, and Haviland (2013) note that in many educational assessments, subscale scores are highly intercorrelated despite earnest attempts by administrators to assess examinee competencies on distinct content domains. Moreover, the total score, which is usually based on many items, is reliable, whereas sub- scale scores often are much less reliable. Bifactor models which include a general ability factor tend to fit student achievement data better than the conventional multidimensional models and the *g*-factors are usually highly reliable whereas the domain specific constructs are much less reliable (Brunner, 2008; Baumert et al., 2009; Saß et al., 2017; Pokropek et al., 2021). These studies are further evidence that the constructs isolated in conventional multi- dimensional models contain a considerable amount of variance attributable to general ability.

1.2. Relationships of factors with covariates

The literature suggests that there are no or only small gender differences among adolescents in general ability (Aluja-Fabregat, Colom, Abad, & Juan-Espinosa, 2000; Colom, Juan-Espinosa, Abad, & Garcia, 2000; Lemos, Abad, Almeida, & Colom, 2013) so gender differences in general ability are not expected. By contrast, the literature suggests that there are stronger gender differences in specific scholastic abilities. In system-wide and international achievement studies, boys typically exhibit higher average test scores in mathematics and girls higher average test scores in reading (Nowell & Hedges, 1998; Harris et al., 2005; Marks, 2008; OECD, 2015). This is also true of estimates of gender gaps in the PISA reading and mathematics tests, albeit with much variation between countries (OECD, 2015; Stoet & Geary, 2015).

In numerous studies, children from lower socioeconomic status (SES) exhibit lower average scores on intelligence tests than their higher SES peers (Bradley & Corwyn, 2002; Schoon, Jones, & Cheng, 2012; Strenze, 2007) a reflection of the importance of social and economic conditions on the development of general ability. A large body of research focuses on the relationship between SES and specific scholastic abilities. The relationship between SES and *g* features in the early childhood development literature but not in the literature on high school students, despite the fact that lower quality educational inputs, both at school and at home, and poorer socio-economic conditions are likely to

reduce children's opportunity to develop general cognitive abilities. Theories contend that SES effects on specific abilities are large because of lower SES students attend lower quality schools, they lack cultural capital, their parents are less interested in their education and provide less support and the effect of these factors weighs more on specific academic abilities. There are many other explanations for SES differences in stu- dent's test scores (see Marks & O'Connell, 2021a, 2021b).

The OECD's ESCS measure incorporates many aspects of the students' environment household possessions, books in the home, home educational resources and the home's cultural resources, it is expected to impact more strongly with general ability and reading than with mathematics and science in which knowledge and skills are taught at school.

Alternatively, SES is strongly associated with general ability among students since parents' cognitive abilities influence their socioeconomic attainments (e.g., education, occupational attainment, income) and their abilities are also transmitted to their children through concerted cultivation efforts which, in turn, influences their performance in ILSAs and other assessments.

Learning time is a key educational resource. According to OECD (2020), learning time has the potential to improve the quality and equity of education outcomes. It is reasonable to expect that the specific learning time students at age 15 spend in the respective domains is positively related to the specific ability factors measured at age 15, for example learning time in mathematics with the mathematics factor, more than it is related to the accumulated general ability factor. The effects of learning time on specific abilities are expected to be stronger than that for general ability.

2. Research questions and analytical strategy

The review of the literature and previous empirical research suggest that, to a considerable extent, PISA might measure general cognitive rather than the domain-specific abilities of reading, mathematics and science. The overarching research question of the present study is to what extent the PISA assessment instruments measures general cognitive ability and domain-specific abilities. This broad research question generates six specific research questions.

(1) Are student responses to the PISA test instruments better represented by the bifactor models than the multidimensional IRT model routinely used in ILSAs?
(2) Are the three dimensions isolated from the multidimensional model conceptually valid or are they simply proxies for general ability?
(3) If bifactor models better represent student responses to PISA test items, which bifactor more closely approximates students' responses: the model with a general ability factor or the model with a general reading factor?
(4) To what extent are PISA scores derived from current scaling procedures are contaminated by a general ability factor?
(5) Are the specific abilities factors generated in the two bifactor models reliable enough to generate student scores?
(6) Are the relationships of gender, learning time and socioeconomic background with the latent ability constructs consistent with theoretical expectations?

To address these research questions, a four-tiered approach is adopted. First, the ability of the three psychometric models discussed— three-dimensional IRT model, the bifactor IRT model and the bifactor (S-1) IRT model—to fit the PISA data are compared using an information criterion (AIC) to compare model fit. Second, the correlations from the multidimensional model are estimated, following Brown (2015) to assess in which countries the three dimensions are sufficient discriminant validity. A variety of bifactor indices are utilized to establish the explanatory power and reliabilities of the latent dimensions generated by the three models. Finally, the statistical relationships between the latent factors derived from the three models and educationally relevant covariates are estimated.

We use this four-tiered approach to provide robust set of evidences that will allow us to answering our questions and minimize the risk of incorrect inferences based on only one approach (cf. Bonifay & Cai, 2017).

3. Materials and methods

3.1. Data

PISA is a low-stake international large-scale assessment administered in both OECD and non-OECD countries. It has been administered to representative samples of 15-year-old students every three years since 2000. The core PISA instruments are the cognitive tests, a student background questionnaire and a school principal questionnaire. In each cycle, there is a major domain for which there are a larger number of items than for the minor domains.

The data analysed in this study are public use files from the 2018 cycle of PISA (downloadable from: http://www.oecd.org/pisa/data). With the application of appropriate weights, the PISA samples are representative of students aged between 15 years and 3 months and 16 years and 2 months at the time of the assessment (generally referred to as 15-year-olds). In each cycle, PISA participants are selected from the population in participating countries through a two-stage random sampling procedure. In the first stage, a stratified sample of schools is drawn, then students are randomly selected from each sampled school. This study is limited to OECD countries because OECD countries make decisions about the development of the PISA instruments and therefore the PISA instruments are primarily, although not solely, designed to respond to the needs of this relatively homogeneous group of countries. Furthermore, in many non-OECD countries student proficiency is considerably lower than in OECD countries and different item sets were administered to better capture the lower proficiency distribution (Rutkowski, Rutkowski, & Liaw, 2018, 2019). Finally, this study is limited to countries that collected achievement data using the computer-based assessment tool because of potential comparability is- sues with the pen and paper approach (Robitzsch, Lüdtke, Goldhammer, Kroehne, & Köller, 2020). All OECD countries implemented the test on computer.

Spain was excluded because of problems with the administration of the reading assessment in 2018 (OECD, 2019c, Annex A9). For Canada, a random sample of 12,994 students was drawn because the full sample (22,653) was too large to perform the computations necessary for this paper using standard hardware.

The data set analysed comprised 226,434 students from 33 OECD countries. The list of countries is presented in the results tables. This sample size applies to the major domain, reading. The sample sizes were considerably smaller for the two minor domains, mathematics and science.

3.2. Measures

In 2018, the PISA test was administered by computer and as in previous cycles a rotation design was used which assigns a subset of items to each student. As in previous PISA cycles, each 2018 PISA assessment was composed of four clusters of domain specific items, with each cluster designed to take around 30 min to complete. Until the 2012 cycle, PISA test items had a fixed position within each cluster and only clusters were rotated across the different assessment forms (OECD, 2012, pp. 29–32). Since 2015, items have been rotated within each cluster.

For the 2018 cycle an adaptive design was implemented for the major domain – reading. This means that the items students are asked to respond to depends on their responses to previous items. Adaptive designs were not implemented for the minor domains. In all domains, the item pool comprised both multiple-choice and constructed response (or open) questions and items varied by format and level of difficulty (OECD, 2019b, Chapter 2).

After completing the assessments, students were administered the background questionnaire which collects data on the education and occupations of their parents, the household's educational and cultural resources and the presence of, and in some instances the number of, a variety of consumer durables. These data are used to create the OECD's standard 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 and Marks & O'Connell, 2021a, 2021b for an extensive critique). The ESCS index is standardised to a mean of zero and a standard deviation of one, across OECD countries. The student questionnaire also collects student demographic data, including gender which is included in these analyses.

The PISA technical report presents, details of data collections, all data cleaning procedures including information on the psychometric characteristics of the items, reliability of scales, and other statistical properties (OECD, 2019b).

3.3. Measurement models

This study analyses the three psychometric models described in Fig. 1 Each model assumes a different latent structure to account for the variation in students' responses to the test items:

3.3.1. *Model 1: three-dimensional IRT model*

This model assumes that three different and correlated latent traits describe students' responses to the test items. The model specifies three specific factors corresponding to the three test domains: reading, math and science. General ability is not included, and the three factors are specified as not independent of each other. This model represents the standard model used in ILSAs.

3.3.2. *Model 2: bifactor IRT model*

The fully symmetrical bifactor model specifies a general ability fac- tor and three specific ability factors. In this model, the domain-specific factors are uncorrelated with the general ability factor and with each other; they comprise only specific factor variance.

3.3.3. *Model 3: bifactor (S-1) IRT model*

Bifactor-(S 1) is a reconfiguration of model 2, where one domain- specific factor (reading) replaces the general factor. In this model, a general reading factor and two domain-specific factors account for variation in item responses. Fig. 1 presents the three models schematically.

All models were estimated using item level data and confirmatory factor analysis (CFA) for categorical data. These models are often referred to as IRT models. IRT and CFA models for categorical data are essentially the same with minor differences in parametrization. The models were estimated by logit link function (or ordered logit link function for partial credit items) and maximum likelihood using the EM algorithm with numerical integration with 15 quadrature points and default convergence criteria settings. Further technical details are available from the first author. All measurement models were estimated using the Mplus version 8 software.

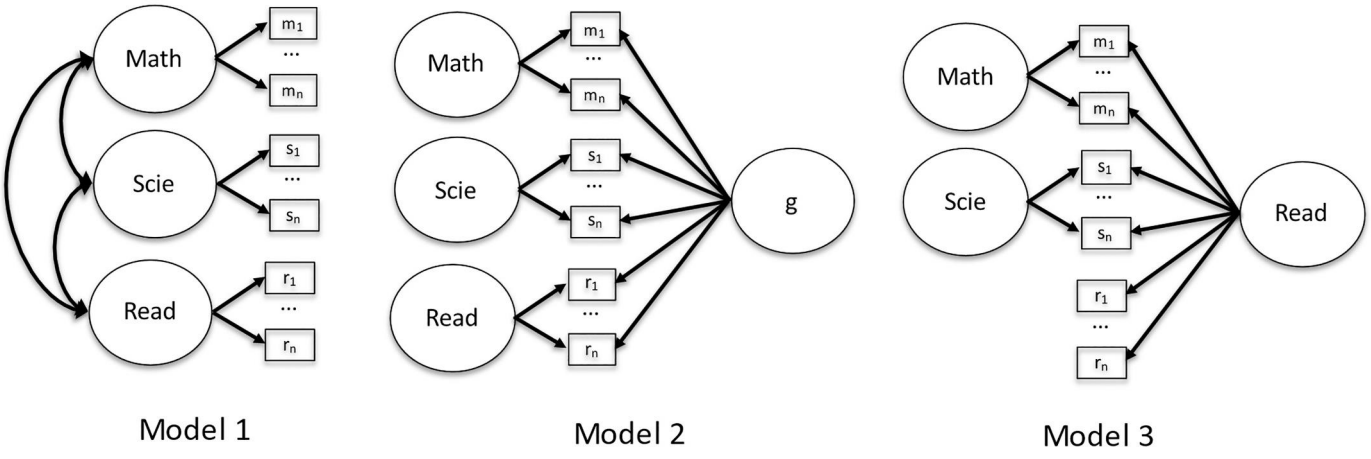

Fig. 1. Alternative models describing response patterns to the PISA test.

3.3.4. *Model fit comparison using an information criterion (AIC)*

The Akaike information criterion (AIC) is used to compare models. This likelihood fit measure penalises free parameters to combat overfitting:

$$AIC = -2 * ln\ (likelihood)\ + 2k, \qquad where\ k\ is\ the\ model\ degrees\ of\ freedom$$

AIC is the preferred fit measure because it is appropriate for PISA data. In contrast, other commonly used measures of model fit—comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), or chi-square based indices—cannot be used because PISA uses an incomplete balanced matrix design in which students answer some, but not all test, questions (as described above). These indices cannot be used to compare models because of the large amount of structured missing data for the cognitive items (Xia and Yang, 2019; Zhang Savalei, 2020; Fitzgerald et al. 2021). Furthermore, the number of students in the assessment (required for BIC) is not known, because in the balanced matrix design, different students answer different numbers of questions. Finally, there is no straightforward way of defining the effective sample size. The complex stratified sampling in PISA means the effective sample size is smaller than simply the number of students tested.

Although there are no strict rules for comparing models, Burnham, Anderson and Huyvaert (2011, p. 25) suggest that differences in AIC around 10 units a strong indication of superior fit. Differences in AIC of around 20 provide

stronger evidence over of a better model. Models with the smallest AIC value are designated as the "better" model, i.e., models whose parameters generates data that more closely approximate the observed data, taking into account differences in the number of free parameters.

3.3.5. Bifactor Model Indices

Various bifactor model indices can assess the validity and reliability of the bifactor model factors (Bonifay, 2020): factor strength, internal consistency reliability, and construct replicability indices. These indices are the Explained Common Variance (ECV) for the general and group factors together with the proportion of uncontaminated correlations (PUP), and the omega family coefficients.

3.3.6. Explained Common Variance (ECV)

The Explained Common Variance (ECV) quantifies the amount of common variance attributable to the general factor and group factors (Rodriguez, Reise, & Haviland, 2016). The ECV is the ratio of the common variance explained by the general factor to the common variance accounted for by the general factor and group factors (Reise, Moore, & Haviland, 2010):

$$ECV = \frac{(\sum \lambda_{Gen}^2)}{(\sum \lambda_{Gen}^2) + (\sum \lambda_{Gr1}^2) + (\sum \lambda_{Gr2}^2) + \cdots + (\sum \lambda_{Grk}^2)} \quad (1)$$

where λ_{Gen}, λ_{Gr1}, λ_{Gr2}, and λ_{Grk} are vectors of standardised factor loadings, with the first term representing the vector of general factor loadings and the other terms – vectors of group factor loadings. Values of ECV range from 0, which indicates no unidimensionality, to 1, indicating a completely unidimensional data structure (Quinn, 2014). High values of ECV suggest that data are essentially unidimensional.

ECV values are largely dependent on the proportion of uncontaminated (by multidimensionality) correlations (PUC), defined as:

$$PUC = \frac{n_{items}(n_{items} - 1)/2 - \Sigma_{group=1}^{group=k} n_{ipg}(n_{ipg} - 1)/2}{n_{items}(n_{items} - 1)/2} \quad (2)$$

Where $n_{items}$is the number of items, and n_{ipg} is the number of items per group for k groups. (Bonifay, Reise, Scheines, & Meijer, 2015:507).

Rodriguez, Reise, and Haviland (2015) suggests that a predominantly unidimensional structure is indicated by ECV values above 0.70 and for PUCs less than 0.70. For these PISA data, the average PUC for the symmetrical bifactor model is 0.64 and 0.90 for the non-symmetrical S-1 bifactor model.

ECV can be computed for group factors, using only a subset of items loading on the specific factor of interest (ECV_{SS}). ECV_{SS} is the proportion of common variance in a subset of items explained by the respective latent specific factor of the common variance accounted for by the general and specific factors:

$$ECV_{SS} = \frac{(\sum \lambda_{Grk}^2)}{(\sum_{i=1}^{j} \lambda_{Gen}^2) + (\sum \lambda_{Grk}^2)} \quad (3)$$

where j denotes the last item for the subset of items loading on the specific factor k. Note that the ECVs do not sum to 1.

Dueber and Toland (2021) show that ECV_{SS} for group factors is useful to determine the strength of a specific factor, that is does it contribute the explained variance for the subset's items beyond that provided by the general factor.

3.3.7. Omega coefficients (ω)

The omega coefficients (ω) measure the amount of reliable variance in unit-weighted composite scores explained by each factor. Coefficient omega is a factor analytic model-based reliability estimate (McDonald, 1999).Coefficient omega is analogous to coefficient alpha in classical measurement theory (Reise, Bonifay, & Haviland, 2013b). Coefficient omega can be calculated for all items, as in equation 4, or for a set of items for a subscale.

$$\omega = \frac{(\sum \lambda_{Gen})^2 + (\sum \lambda_{Gr1})^2 + (\sum \lambda_{Gr2})^2 + \cdots + (\sum \lambda_{Grk})^2}{(\sum \lambda_{Gen})^2 + (\sum \lambda_{Gr1})^2 + (\sum \lambda_{Gr2})^2 + \cdots + (\sum \lambda_{Grk})^2 + \Sigma(1-h)^2} \quad (4)$$

Omega hierarchical (ωH) quantifies the proportion of reliable variance in students' responses due to the general factor:

$$\omega_H = \frac{(\sum \lambda_{Gen})^2}{(\sum \lambda_{Gen})^2 + (\sum \lambda_{Gr1})^2 + (\sum \lambda_{Gr2})^2 + \cdots + (\sum \lambda_{Grk})^2 + \Sigma(1-h)^2} \quad (3)$$

where λs represent vectors of standardised factor loadings and $\Sigma(1-h)^2$ is the sum of the items' unique variances. Values of ωH above 0.8 indicate that the unit-weighted total scores are essentially unidimensional (Rodriguez et al., 2015).

OmegaS is an estimate of composite reliability of subscores and can be computed for each specific factor:

$$\omega_S = \frac{(\sum_{i=1}^{j} \lambda_{Gen})^2 + (\sum \lambda_{Grk})^2}{(\sum_{i=1}^{j} \lambda_{Gen})^2 + (\sum \lambda_{Grk})^2 + \Sigma(1-h)^2} \quad (4)$$

Note that the general factor contributes to the reliability of the specific factor.

Omega hierarchical subscale (ωHS) quantifies the degree that subscale scores (for reading, mathematics, and science) are (not) confounded by the general factor (Reise, Bonifay, & Haviland, 2013b). It is calculated similarly to omega hierarchical, but just for a subset of items:

$$\omega_{HS} = \frac{(\sum \lambda_{Grk})^2}{(\sum_{i=1}^{j} \lambda_{Gen})^2 + (\sum \lambda_{Grk})^2 + \Sigma(1-h)^2} \quad (5)$$

Dueber and Toland (2021) show that for a subset of items, Omega hierarchical (ω_{HS}) and omega subscale (ω_S) can be used to determine if subscores were generated, would they add value beyond that provided by the total score? However, Dueber and Toland (2021) stress that once a specific factor is identified as adding value it does not automatically mean that scores generated from the specific factor can be usefully interpreted. If subscores are to be interpreted, a high omega subscale "can be considered as evidence that such an interpretation is statistically appropriate" (Dueber & Toland, 2021, p.15). In other words, a signal is observable but the signal to noise ratio is too low for it to be reliably measured. The cut-off values proposed by Dueber and Toland (2021) for assessing if specific factors in the SEM framework and the subscales add value are utilized in the results section and can be found in the Supplementary online Annex.

3.4. Nomological Network (Covariate Analyses with Structural Equation Modelling)

Structural Equation Models (SEM) comprising of measurement and structural models are used to estimate the relationships between the set of covariates and the latent abilities. The measurement models are the three factor analytic models discussed above. The covariates are learning time in language of the test, learning time in mathematics and science, gender and socio-economic status. Standardised coefficients were estimated for 32 countries and summarised by boxplots. The coefficients are standardised regression coefficients: the predicted average change in the dependent variable (here, always latent abilities) for a one-standard increase in the respective predictor variable with all other variables held constant. Missing data was handled through full information maximum likelihood estimation. Japan is not included in the analyses of the relationships between latent dimensions of covariates because many of the measures were not available.

4. Results

4.1. Model fit comparison using an information criterion (AIC)

Table 1 presents AIC values for the three-dimensional model, the symmetrical bifactor model, and the S-1 reading model. The table in- cludes differences in AIC values between the three models and the correlations between factors in the conventional three-dimensional model. The AIC values indicate that the symmetrical bifactor model outperforms the two other models in all countries and by a considerable margin in the great majority of countries. Differences in AIC between the symmetrical bifactor and the three-dimensional model are larger than 100 in 30 out of the 33 countries (Iceland, the Netherlands and the United States are the exceptions).

AIC differences between the symmetrical bifactor and the S-1 reading models are larger than 150 in 32 out of the 33 countries (the United States, the sole exception with a difference of 110). If for technical reasons, bifactor models fit the data better, then the reading model would produce similar fits as the general ability bifactor model. The results are unequivocable, the bifactor provides the best fit to the data in all 33 countries, this is despite AIC penalizing models with more degrees of freedom.

Model Comparisons using AIC criterion and correlations of latent variables from 3D model.

		AIC			ΔAIC			Correlations 3D Model		
Country	N	3D	Bif	S-1	3D -Bif	3D -S-1	Bif - S-1	Math with Reading	Math with Science	Reading with Science
Australia	14,273	933,042	932,109	932,962	933	81	−852	0.834	0.916	0.899
Austria	6801	444,978	444,507	444,845	471	133	−338	0.867	0.932	0.917
Belgium	8474	536,879	536,641	536,902	239	−22	−261	0.867	0.943	0.919
Canada	12,994	747,032	746,084	746,799	948	233	−715	0.812	0.898	0.900
Czech Rep.	7019	453,258	452,734	453,196	524	62	−462	0.857	0.892	0.906
Denmark	7655	491,267	490,919	491,248	349	20	−329	0.832	0.903	0.884
Estonia	5316	353,299	353,100	353,281	200	18	−182	0.829	0.894	0.900
Finland	5648	356,282	356,017	356,296	266	−13	−279	0.835	0.899	0.901
France	6306	394,678	394,348	394,691	330	−13	−343	0.883	0.904	0.92
Germany	5451	340,869	340,721	340,923	148	−54	−202	0.873	0.921	0.911
Greece	6398	366,448	365,863	366,491	585	−43	−628	0.827	0.794	0.894
Hungary	5130	337,300	337,005	337,299	296	1	−295	0.876	0.912	0.913
Iceland	3296	207,205	207,173	207,348	32	−143	−175	0.803	0.935	0.889
Ireland	5577	369,314	369,179	369,341	135	−27	−162	0.852	0.901	0.903
Israel	6618	352,652	352,246	352,675	406	−23	−429	0.862	0.905	0.921
Italy	11,784	765,571	764,239	765,417	1332	153	−1178	0.817	0.891	0.879
Japan	6108	394,920	394,612	395,022	308	−102	−409	0.822	0.915	0.901
Korea	6648	385,136	384,963	385,223	173	−87	−260	0.843	0.927	0.888
Latvia	5298	319,269	319,089	319,336	180	−67	−247	0.825	0.898	0.881
Lithuania	6883	405,659	405,423	405,693	236	−34	−270	0.848	0.853	0.909
Luxembourg	5230	327,757	327,491	327,782	267	−24	−291	0.862	0.851	0.920
Netherlands	4765	293,172	293,115	293,298	57	−126	−184	0.864	0.93	0.905
New Zealand	6173	401,084	400,847	401,039	238	46	−192	0.821	0.846	0.911
Norway	5810	364,799	364,560	364,830	239	−31	−270	0.835	0.92	0.877
Poland	5624	372,516	371,966	372,535	550	−18	−568	0.822	0.845	0.887
Portugal	5932	383,314	383,202	383,380	112	−67	−178	0.854	0.896	0.890
Slovak Rep.	5962	346,204	345,684	346,259	519	−55	−575	0.853	0.859	0.895
Slovenia	6401	425,931	425,743	426,037	188	−106	−294	0.812	0.952	0.873
Sweden	5503	332,130	331,898	332,140	232	−10	−242	0.84	0.952	0.895
Switzerland	5822	380,886	380,617	380,922	269	−37	−305	0.85	0.914	0.897
Turkey	6890	460,086	459,703	460,152	383	−66	−449	0.856	0.918	0.931
UK	13,807	900,534	899,671	900,554	863	−20	−882	0.837	0.904	0.894
United States	4838	316,430	316,377	316,487	53	−57	−110	0.872	0.927	0.911

Note: 3D – model with three dimensions correlated, Bif – symmetric bifactor model, S-1 – bifactor (S-1) model; ΔAIC – differences in AIC values between respective models. Smallest AIC value is selected as the "best" model.

4.2. *Factor correlations from the correlated-factor model*

Table 1 presents the correlations between the reading, mathematics and science factors generated by the multidimensional model. The correlations between the reading and mathematics factors are above 0.8 in all 33 countries. For mathematics and science, the correlations are generally higher sometimes greater than 0.90. For reading and science the within country correlations are generally between the other two correlations. Since all but one of the between-factor correlations are above 0.8, the three factors in the multidimensional model exhibit poor discriminant validity, so are not viable as measures of distinct concepts.

4.3. *Bifactor model indices*

Table 2 presents four indices for the symmetrical bifactor model: the ECVs, omega coefficient for the general and specific factors, omega hierarchical for the general factor, and the omega subscale coefficients for the specific ability

factors.

For the symmetric bifactor model, the ECV values are high, around 0.8 or more. This means that the general factor accounts for 80% or more of the common variance in item responses. Therefore, a simpler unidimensional model would fit the data nearly as well the symmetric bifactor model. The amount of common variance in the subdomain items explained by its respective specific factor varies, with an average of 0.27 for mathematics, 0.17 for science and to 0.12 for reading. So most of the common variance in the subdomain items is accounted for by the gen- eral factor, not the respective specific factors. According to Dueber and Toland (2021) cut-offs, there is no added value in specifying an additional specific ability in the SEM framework: for mathematics in Korea and Slovenia; for reading in Australia, Canada, Czech Republic, Finland, France, Germany, Israel, Lithuania, Luxembourg, the Netherlands, New Zealand, Norway, Sweden, and the United Kingdom; for science in Belgium, Denmark, Korea, Slovenia, the United States, and Turkey.

For the rest of the countries, the subdomain specific ECVs for reading and science, even though they are above the cut-off values proposed by Dueber and Toland (2021), are still very low, often below 0.2, that is they account for less than 20% of common variance in the subdomains. In other words, more than 80% of the common variance for the reading and science items is attributable to the general factor.

The omega coefficients suggest that a large part of the unit-weighted total score variance is due to the general ability factor and the subscale scores are highly confounded by this factor in all countries. All the ω_{H} values exceed 0.8 which indicates that the unit-weighted total scores are essentially unidimensional (Rodriguez et al., 2015).

For reading, on average, only 5% of the reliable unit-weighted variance is due to the specific factor (0.044/0.982). The remainder is due to the general factor. For mathematics, on average 24% of unit- weighted reliable variance in the mathematics items can be attributed a specific mathematics factor (0.221/0.938). For Belgium, Denmark, Korea, Portugal, Slovenia, and the United States proportions are considerably lower. For science, on average 14% of unit-weighted reliable variance in the science items can be attributed a specific science factor (0. 135/0.952). For Belgium, Denmark, Korea, Portugal, Slovenia, Turkey, and the United States the specific science factors account for 6. Overall results presented in Table 2 indicate that subscale scores for reading and science would not add value beyond scores generated from

Table 2
Bifactor indices for general and specific factors (symmetric bifactor model).

	ECV				Omega				Omega hierarchical & subscale			
Country	*g*	Read	Math	Science	*g*	Read	Math	Science	*g*	Read	Math	Science
Australia	0.858	0.080	0.314	0.182	0.989	0.984	0.939	0.956	0.962	0.012	0.264	0.171
Austria	0.855	0.100	0.272	0.166	0.988	0.982	0.939	0.954	0.965	0.013	0.208	0.146
Belgium	0.855	0.175	0.164	0.062	0.988	0.981	0.941	0.954	0.932	0.143	0.119	0.002
Canada	0.846	0.085	0.350	0.190	0.987	0.982	0.939	0.951	0.961	0.010	0.293	0.170
Czech Republic	0.858	0.089	0.274	0.184	0.988	0.982	0.944	0.956	0.964	0.006	0.237	0.153
Denmark	0.811	0.232	0.198	0.089	0.987	0.980	0.932	0.952	0.911	0.201	0.123	0.016
Estonia	0.834	0.115	0.306	0.192	0.985	0.979	0.934	0.946	0.953	0.023	0.258	0.151
Finland	0.847	0.096	0.315	0.195	0.987	0.982	0.933	0.953	0.962	0.005	0.275	0.173
France	0.870	0.088	0.235	0.162	0.989	0.984	0.946	0.955	0.973	0.000	0.187	0.143
Germany	0.859	0.099	0.246	0.171	0.989	0.983	0.943	0.958	0.965	0.014	0.203	0.153
Greece	0.814	0.130	0.361	0.209	0.986	0.981	0.930	0.941	0.964	0.003	0.286	0.173
Hungary	0.861	0.101	0.244	0.159	0.988	0.981	0.936	0.952	0.965	0.015	0.202	0.117
Iceland	0.821	0.120	0.345	0.222	0.988	0.983	0.936	0.951	0.960	0.012	0.307	0.191
Ireland	0.839	0.116	0.294	0.185	0.985	0.979	0.920	0.948	0.961	0.017	0.220	0.155
Israel	0.870	0.070	0.295	0.166	0.991	0.987	0.951	0.962	0.971	0.005	0.236	0.148
Italy	0.820	0.114	0.338	0.229	0.986	0.980	0.934	0.947	0.951	0.025	0.287	0.202
Japan	0.825	0.126	0.320	0.196	0.986	0.980	0.937	0.951	0.959	0.016	0.280	0.163
Korea	0.803	0.278	0.099	0.080	0.988	0.982	0.943	0.955	0.906	0.239	0.025	0.012
Latvia	0.815	0.124	0.342	0.221	0.985	0.978	0.926	0.943	0.956	0.014	0.271	0.182
Lithuania	0.854	0.084	0.298	0.189	0.987	0.981	0.937	0.949	0.964	0.007	0.258	0.163
Luxembourg	0.862	0.093	0.269	0.165	0.989	0.985	0.942	0.956	0.967	0.016	0.218	0.145
Netherlands	0.860	0.098	0.242	0.178	0.989	0.984	0.943	0.960	0.963	0.023	0.205	0.159
New Zealand	0.851	0.089	0.340	0.169	0.989	0.984	0.941	0.957	0.960	0.022	0.286	0.154
Norway	0.842	0.090	0.305	0.235	0.988	0.983	0.937	0.954	0.962	0.000	0.271	0.208
Poland	0.823	0.118	0.328	0.220	0.986	0.980	0.939	0.952	0.957	0.006	0.278	0.196
Portugal	0.833	0.174	0.206	0.119	0.988	0.982	0.943	0.953	0.932	0.127	0.162	0.060
Slovak Republic	0.836	0.117	0.284	0.197	0.988	0.983	0.942	0.957	0.962	0.015	0.234	0.170
Slovenia	0.794	0.282	0.139	0.076	0.987	0.981	0.933	0.949	0.893	0.250	0.046	0.006
Sweden	0.843	0.097	0.312	0.206	0.988	0.983	0.939	0.953	0.962	0.011	0.260	0.178
Switzerland	0.844	0.109	0.273	0.191	0.988	0.982	0.940	0.954	0.960	0.019	0.231	0.167
Turkey	0.855	0.109	0.280	0.118	0.986	0.978	0.937	0.934	0.971	0.014	0.233	0.026
United Kingdom	0.850	0.092	0.292	0.194	0.987	0.981	0.936	0.951	0.957	0.020	0.258	0.178
United States	0.843	0.185	0.173	0.074	0.990	0.985	0.943	0.958	0.926	0.158	0.084	0.011
Average	0.841	0.124	0.274	0.169	0.988	0.982	0.938	0.952	0.954	0.044	0.221	0.135

the general factor because in most countries, the signal to noise ratio is too low. Although the psychometric properties of the mathematics factors are somewhat better, a reliable variance of between 20 and 30% of is not sufficient to provide valid inferences about science scores generated from the latent mathematics factor. However, because there is some value-added information beyond what is conveyed by the general ability factor a psychometric model applied to other data could be developed to generate viable mathematics subscores that are independent of scores generated from the general factor.

Table 3 presents bifactor indices for the S-1 reading model where reading is the general factor. The ECV values are very high, higher than those for the symmetrical bifactor model. The indices indicate that in almost all countries a latent reading factor explains around 90% of common variance, is responsible for almost all of the reliable variance in unit-weighted total scores, and that the remaining group factors (math and science) are much weaker. Note that reading factor incorporated much of the variance attributed to general cognitive ability in the pre- vious bifactor model.

On average, the domain specific ECV for mathematics is 0.31, and 0.20 for science. The omega subscale estimates are 0.26 for mathematics and 0.19 for science. The mathematics and science factors are slightly more reliable in the S-1 reading model than the previous bifactor model, but not reliable enough to generate valid subset scores.

4.4. Nomological network (covariate analyses with structural equation modelling)

Fig. 2 presents results from structural equation models which estimate the relationships between the covariates (gender, socio-economic characteristics and learning time) and the latent variables isolated

Table 3
Bifactor indices for general and specific factors (bifactor S-1 model). from the three models.

The box and whisker plots summarise country specific associations and illustrate the variability in associations across countries. The width of the box is the interquartile range bounded by the upper and lower quartiles. The median is the vertical line inside the box. The value that the whisker begins is the lower quartile minus 1.5 times the interquartile range. The value that whisker ends is the upper quartile plus 1.5 times the interquartile range. The dots represent outlier countries outside the range of the whiskers. All coefficients in Fig. 2 and in the more detailed tables in the Supplementary online Annex are standardised.

In the bifactor model, girls score, on average, higher than boys on the general ability factor, although there is much variation between countries (Fig. 2). In the two bifactor models, boys, on average, score higher than girls in mathematics and science. These gender differences are considerably larger than those from the multidimensional model. Similarly, gender differences in the multidimensional model favouring girls in reading are smaller than in the bifactor models. A plausible interpretation of these results is that the multidimensional model un- derestimates gender gaps in reading (favouring girls), and in mathematics and science (favouring boys).

The multidimensional model shows plausible relationships with ESCS. Contrary, to expectations, ESCS effects are not stronger for reading than for mathematics and science; the average estimate is around 0.26. In the bifactor general ability model, the ESCS effect is also around 0.25. According to the general ability bifactor model, ESCS is not significantly associated with the reading and science factors in most countries and only weakly associated with mathematics. According to the bifactor reading model, ESCS is moderately correlated with reading, again at around 0.26, but only weakly with mathematics and science. This pattern is contrary to theoretical expectations but is explicable if it

	ECV			Omega			Omega hierarchical & subscale		
Country	Read	Math	Science	Read	Math	Science	Read	Math	Science
Australia	0.900	0.324	0.195	0.989	0.939	0.956	0.965	0.276	0.184
Austria	0.908	0.282	0.176	0.988	0.939	0.954	0.968	0.219	0.157
Belgium	0.913	0.259	0.161	0.988	0.942	0.953	0.968	0.216	0.141
Canada	0.890	0.361	0.200	0.987	0.938	0.951	0.962	0.304	0.180
Czech Republic	0.906	0.279	0.190	0.988	0.944	0.956	0.966	0.242	0.167
Denmark	0.887	0.327	0.220	0.987	0.932	0.952	0.962	0.272	0.197
Estonia	0.893	0.322	0.204	0.985	0.934	0.944	0.960	0.275	0.168
Finland	0.899	0.316	0.202	0.987	0.933	0.953	0.963	0.277	0.181
France	0.916	0.243	0.171	0.989	0.946	0.955	0.972	0.195	0.154
Germany	0.910	0.258	0.181	0.989	0.943	0.958	0.968	0.216	0.165
Greece	0.887	0.367	0.217	0.986	0.930	0.941	0.965	0.296	0.185
Hungary	0.910	0.256	0.177	0.987	0.936	0.952	0.969	0.216	0.141
Iceland	0.888	0.354	0.223	0.988	0.937	0.951	0.964	0.317	0.193
Ireland	0.899	0.306	0.201	0.985	0.920	0.946	0.963	0.235	0.174
Israel	0.909	0.297	0.167	0.991	0.950	0.962	0.973	0.239	0.149
Italy	0.879	0.349	0.239	0.986	0.935	0.947	0.958	0.298	0.213
Japan	0.892	0.329	0.202	0.986	0.937	0.951	0.962	0.290	0.171
Korea	0.896	0.297	0.215	0.988	0.945	0.955	0.961	0.262	0.186
Latvia	0.879	0.356	0.232	0.985	0.926	0.943	0.959	0.288	0.197
Lithuania	0.899	0.301	0.192	0.987	0.937	0.949	0.966	0.261	0.166
Luxembourg	0.913	0.277	0.169	0.989	0.942	0.956	0.971	0.227	0.149
Netherlands	0.909	0.262	0.187	0.989	0.942	0.960	0.969	0.216	0.170
New Zealand	0.896	0.352	0.181	0.989	0.941	0.957	0.965	0.298	0.168
Norway	0.890	0.308	0.238	0.988	0.937	0.954	0.962	0.273	0.212
Poland	0.885	0.332	0.226	0.986	0.939	0.952	0.958	0.283	0.202
Portugal	0.898	0.278	0.220	0.988	0.944	0.953	0.964	0.237	0.196
Slovak Republic	0.895	0.298	0.209	0.988	0.942	0.957	0.965	0.249	0.185
Slovenia	0.883	0.340	0.239	0.986	0.933	0.950	0.959	0.297	0.219
Sweden	0.896	0.315	0.210	0.988	0.939	0.953	0.965	0.265	0.184
Switzerland	0.900	0.285	0.205	0.988	0.940	0.954	0.965	0.244	0.183
Turkey	0.908	0.284	0.144	0.985	0.937	0.936	0.968	0.239	0.107
United Kingdom	0.897	0.305	0.205	0.987	0.936	0.951	0.962	0.271	0.190
United States	0.908	0.283	0.182	0.990	0.942	0.959	0.970	0.225	0.164
Average	0.898	0.306	0.199	0.988	0.938	0.952	0.965	0.258	0.176

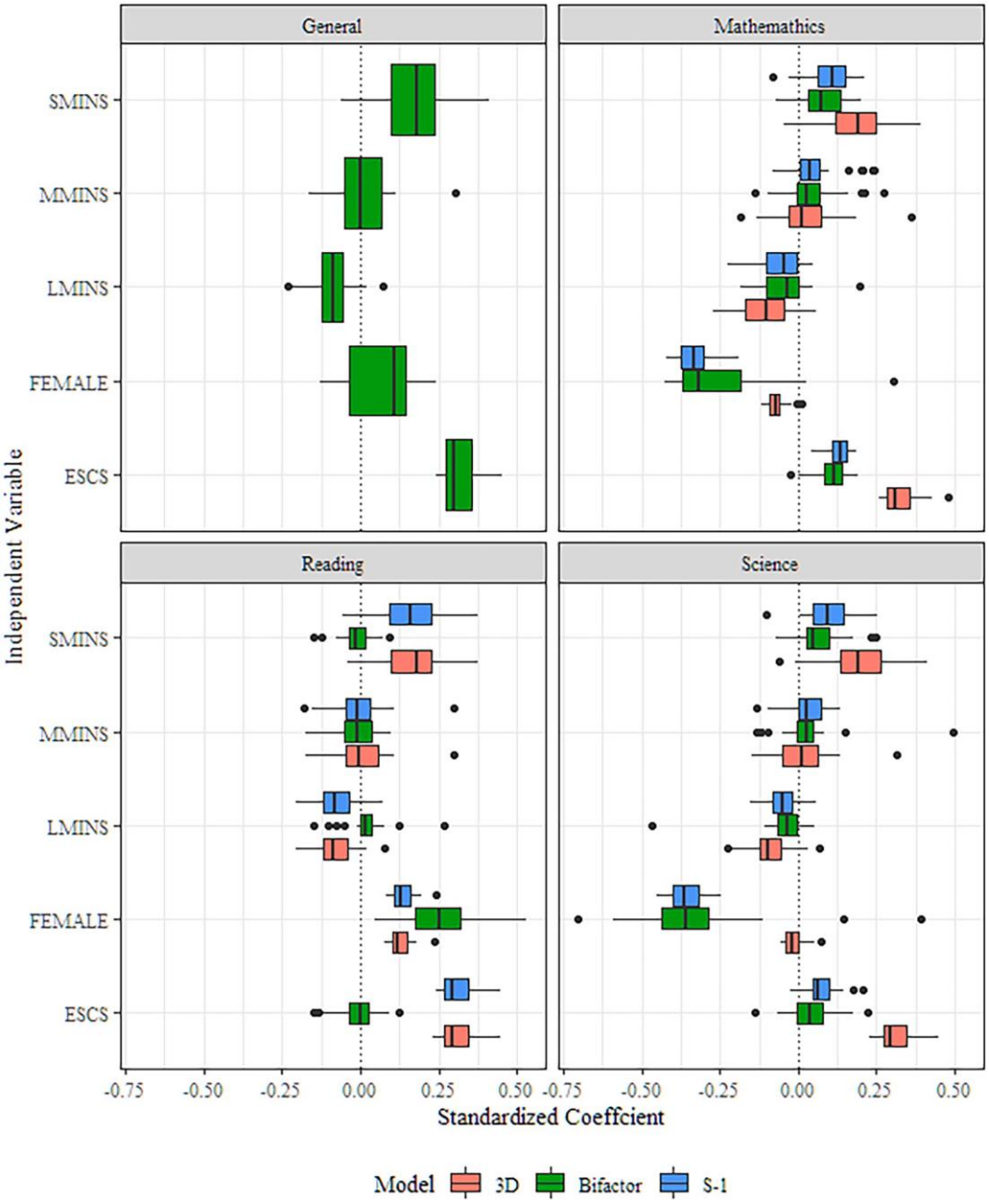

Fig. 2. Standardised regression coefficients from SEM models where abilities were predicted by a set of explanatory variables.
Note: SMINS - science learning time; MMINS - mathematics learning time; LMINS - language learning time; FEMALE – effect for females (boys as reference); ESCS - PISA measure of socioeconomic status.

is assumed that the reading factor in the S1 bifactor model incorporates general cognitive ability.

The bifactor general ability model provides most plausible explana- tion for these associations. The estimates are much the same for the three domains because they are proxy measures of cognitive ability. ESCS is associated with general cognitive ability because parents' socioeconomic attainments are associated with their general ability which is trans- mitted to their children.

The estimates from the analyses of learning time do not conform to theoretical expectations. Learning time in the test-language lessons has mostly negative relationships with the reading factor in the multi- dimensional and reading bifactor models. In the bifactor model, its effects are generally positive but very small. Contrary to expectations, in the three dimensional and the S-1 reading models, learning time in

science is more strongly associated with reading than learning time in the language-of-test. In the bifactor model, learning time in science has no relationship with reading.

In most countries, learning time in mathematics is only very weakly associated with the mathematics factors according to all three models. The average effect for learning time in mathematics with the mathematics factor is close to zero. Similarly learning time in science is weakly associated with the science factor. Learning time for science has similar effects for science and mathematics factors.

The lack of correspond between learning time in a domain and the respective may be because learning time is a poor measure of students' knowledge and skills. In addition, learning time may have non-linear relationships with student performance, both high and low performers have more learning time but for different reasons.

5. Conclusions and discussion

This study compared the standard multidimensional model typically used in ILSAs studies, with the bifactor model specifying a general ability factor, and the S-1 reading bifactor model where the general factor is reading rather than general ability. The data analysed were students' responses to the 2018 PISA instruments tests in 33 OECD countries. Model fits were appropriately compared by the Akaike in- formation criterion which adjusts for the number of free parameters. For the multidimensional model, the extent that the three dimensions exhibited discriminant validity was assessed by the correlations of the three factors. The explanatory power and reliabilities of the latent factors isolated from the two bifactor models were assessed by a range of bifactor indices. Finally, gender and socioeconomic differences in the latent factors isolated from the models were compared as well the impact of learning time.

The overall conclusion of this study is that the bifactor model incorporating a general ability factor yields the best representation of students' responses in the PISA 2018 test. Education specialists might refer to this factor as general academic ability whereas psychologists might refer to this factor as intelligence. Our work simply allows us to identify a general factor that explains, to a large extent, individual variations in achievement tasks in different domains but, in the absence of a pre-defined framework incorporating such construct, does not allow us to define it. The interpretation of this common factor differs depending on the level at which it is considered. At the individual level, this general factor reflects personal experiences and innate potential. At the population level it reflects the cumulate effects of the social and economic context children encounter as well as the overall quality of the education systems in which they grow and develop. The bifactor model. Considerably outperforms the multi-dimensional model. In most countries considered, the differences in the AIC fit index are very large. The bifactor general ability model also exhibits superior fit indices to the bifactor reading model indicating that the superior fit is substantively important and is not a statistical artefact. These results were consistent across the 33 countries. The symmetrical bifactor general ability model shows that 80%, or more, of the common variance in student responses to the PISA 2018 instruments is accounted for by a general ability factor. This percentage is remarkably consistent across countries. Only the mathematics factor makes a non-trivial contribution to the explained variance, beyond that of the general cognitive ability. On average, 27% of variance in the mathematics items is independent of the general factor and can be attributed to a specific mathematics ability factor. The respective estimates for reading and science are 12% and 17%. How- ever, the reliabilities of the mathematics factor (when considering g) are too low to generate interpretable mathematics scale scores.

Results indicating that the relationship between learning time and abilities measures derived using the bifactor models are weak could be an indication that the currently PISA test instruments may not contain enough information to measure specific educational abilities. One would in fact expect learning time in a specific subject, for example mathematics, to be associated with how well a student achieves in mathematics, net of general and other abilities. At the same time, although at the theoretical level the relationship between domain specific abilities and learning time should reveal meaningful associations, it might be difficult to observe such associations empirically. Factors such as varying quality of teachers' instructional practices, differences in the organisation of the curriculum, in the timetable and the school year, and students' motivation to learn could all mean that the intended time devoted to learning is not effectively translated in effective time used to learn (Scheerens & Hendriks, 2014). Moreover, even countries' level of economic development can mediate or condition how effectively learning time translates into the acquisition of

domain specific abilities (Baker et al., 2004). Future studies could try to overcome these short- comings by linking the PISA specific-domain factors to external measures of student performance (grades, marks, test scores) in the respective subject areas to establish differential predictive validity. In fact, this would be an important extension to our study.

This work suggests that the relationships between PISA scores generated from the multidimensional model with explanatory variables such as gender and socio-economic status reflect to a large extent gen- eral ability rather than domain specific abilities. In light of this work, analysts should be mindful about the possible contamination of the specific domains with general ability when considering relationships between PISA test scores referring to specific domains and socioeconomic, demographic, school, and teacher variables.

PISA is complex, ambitious and innovative assessment that, over the years, has evolved greatly to reflect advances in assessment methodologies, analytical techniques, administration possibilities, and computing power. Research and development is a key component of the PISA programme and investments in research and development have allowed PISA to evolve and experiment over the years. Examples include the rotation of background questionnaire materials in PISA 2012, the shift to computer-based administration from 2015, the use of adaptive testing since 2018. While several innovations have been incorporated in successive editions of PISA, thus far, item development and selection in PISA have not considered theoretical models that reflect the relevance of general ability and domain specific abilities and, as such, the domain specific factors estimated with the bifactor model have low reliability.

Our work suggests that analyses conducted ex post can identify problems and inconsistencies in estimates when different scaling models are used. At the same time, analyses conducted ex post cannot alter the nature of the data. By contrast, alternative measurement frameworks could allow to better characterise the respective roles of general and specific abilities and to guide policy. On top of the core domains of reading, mathematics and science, since 2012, PISA has pioneered the development of new assessment domains. These include domain general problem solving and financial literacy in PISA 2012, collaborative problem solving in PISA 2015, global competence in PISA 2018 and creativity in PISA 2022. Our work, indicating the importance of considering the role of general ability in determining students' results on the PISA test appears to be even more critical given these developments in the nature of the PISA test. In response to our work, investments in the development of assessment frameworks that reflect a theoretical model that incorporates the role of general ability should be considered. Such a framework could then guide the development of assessment items that allow to measure domain specific abilities after considering the role of general ability with greater reliability and possibly derive sounder pol- icy implications. The bifactor model should be incorporated into both item selection and analysis of LSAs. Such an approach would allow re- searchers to disentangle general abilities from subject specific abilities and enable the testing hypotheses of influences on specific abilities.

The results reported in this work do not advocate for a generalised endorsement of one model across all large-scale assessments but, rather, that when developing assessments, careful consideration should be given to the potential role of general ability and that appropriate items should be selected to account for the role of general ability in shaping test results. The aim of such exercise would be to ensure that domain- specific abilities that assessments intend to measure can be effectively measured, given the instruments being developed and fielded. In other words, it is recognised that the importance of general ability may and should vary across ILSAs. In assessments such as TIMSS where students are asked to demonstrate their ability to apply their skills learnt at school to specific problems, general ability should be less important than in problem-based assessments such as PISA. Correctly answering questions on trigonometry, solving quadratic equations, calculus, physics and chemistry, requires using specific skills learnt at school taught by mathematics or science teachers rather than general ability throughout one.

Proponents of the bifactor S-1 reading model stress that this model is easier to interpret than the bifactor model since the general factor in the S-1 reading model is defined by the omitted (reference) factor (Eid et al., 2018). However, the S-1 reading model may be an example of the naming fallacy (Kline, 2016). It cannot be assumed that the name of factor defines what it actually is. A more accurate designation would be the general ability-plus-reading factor model, since the model combines general cognitive ability with reading, which accounts for its very high reliabilities. As mentioned above, the general ability-plus-reading factor model fits the data less well than the symmetrical bifactor in all 33 countries examined.

The five research literatures discussed in the introduction are a reminder for researchers and education policy makers that the empirical findings presented in this work are not merely a statistical exercise but, rather, support prior theoretical work on the importance of general ability for student performance in ILSAs and other achievement tests. In the past, social science research in general, and education policy research in particular, may have neglected the importance of general ability in explaining student outcomes because of the widespread - yet wrong notion - that recognising and emphasising the role of differences in general ability would weaken policy commitments to ensure social justice and equality in and through education. Despite considerable in- vestments, educational outcomes remain highly unequal throughout the world. Ignoring the role of general ability in shaping the variation in student outcomes will not help create a fair society, because it will not contribute to derive sound policy implications and the best evidence- based policy interventions.

Acknowledgements

Francesca Borgonovi acknowledges support from the British Academy through its Global Professorship scheme. The views expressed in this piece are of the authors and do not necessarily reflect those of the partner institutions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi. org/10.1016/j.intell.2022.101653

References

Aluja-Fabregat, A., Colom, R., Abad, F., & Juan-Espinosa, M. (2000). Sex differences in general intelligence defined as g among young adolescents. Personality and Individual Differences, 28(4), 813-820.

Andersen, S. C., Humlum, M. K., & Nandrup, A. B. (2016). Increasing instruction time in school does increase learning. Proceedings of the National Academy of Sciences, 113(27), 7481-7484.

Arias, V. B., Ponce, F. P., & Núñez, D. E. (2018). Bifactor Models of Attention-Deficit/Hyperactivity Disorder (ADHD): An Evaluation of Three Necessary but Underused Psychometric Indexes. Assessment, 25(7), 885–897. https://doi.org/10.1177/1073191116679260

Armor, D. J. (2003). Maximizing intelligence. New Brunswick and London: Transaction Publishers.

Asbury, K., & Plomin, R. (2014). G is for genes: The impact of genetics on education and achievement. Chichester, West Sussex: Wiley-Blackwell.

Baker. et al. (2004), "Instructional Time and National Achievement: Cross-National Evidence", PROSPECTS, Vol. 34/3, pp. 311-334, http://dx.doi.org/10.1007/s11125-004-5310-1.

Barnard-Brak, L., Stevens, T., & Ritter, W. (2017). Reading and mathematics equally important to science achievement: Results from nationally-representative data. Learning and Individual Differences, 58, 1-9. https://doi.org/10.1016/j.lindif.2017.07.001

Bartels, M., Rietveld, M. J. H., Van Baal, G. C. M., & Boomsma, D. I. (2002). Heritability of educational achievement in 12-year-olds and the overlap with cognitive ability. Twin Research, 5(6), 544-553. https://doi.org/10.1375/136905202762342017

Baumert, J., Nagy, G., & Lehmann, R. (2012). Cumulative advantages and the emergence of social and ethnic inequality: Matthew effects in reading and mathematics development within elementary schools? Child Development, 83(4), 1347-1367. https://doi.org/10.1111/j.1467-8624.2012.01779.x.

Bentler, P. M. (2009). Alpha, Dimension-Free, and Model-Based Internal Consistency Reliability. Psychometrika, 74(1), 137–143. https://doi.org/10.1007/S11336-008-9100-1

Berge, J. M. F. ten, & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. Psychometrika, 69(4), 613–625. https://doi.org/10.1007/BF02289858

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the social sciences. Mahwah, NJ: Lawrence Erlbaum Associates.

Bonifay, W. (2020). Multidimensional item response theory. Quantitative applications in the social sciences: Vol. 183. Sage.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When Are Multidimensional Data Unidimensional Enough for Structural Equation Modeling? An Evaluation of the DETECT Multidimensionality Index. Structural Equation Modeling: A Multidisciplinary Journal, 22(4), 504–516. https://doi.org/10.1080/10705511.2014.938596

Bonifay, W., & Cai, L. (2017). On the complexity of Item Response Theory models. Multivariate Behavioral Research, 52(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Box, George E. P. (1976) Science and Statistics, Journal of the American Statistical Association, 71:356, 791-799, DOI: 10.1080/01621459.1976.10480949

Bradley, R.H. & Corwyn R.F. (2002). Socioeconomic status & child development. Annual Review of Psychology, 53, 371-399.

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. Paris: OECD Publishing http://dx.doi.org/10.1787/5k9fdfqffr28-en

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). The Guilford Press.

Burgaleta, M., Head, K., Álvarez-Linera, J., Martínez, K., Escorial, S., Haier, R., & Colom, R. (2012). Sex differences in brain volume are related to specific skills, not to general intelligence. Intelligence, 40(1), 60-68.

Calvin, C. M., Fernandes, C., Smith, P., Visscher, P. M., & Deary, I. J. (2010). Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in England. Intelligence, 38(4), 424-432. https://doi.org/10.1016/j.intell.2010.04.005

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. Studies in educational evaluation, 49, 30-41.

Colom, R., Juan-Espinosa, M., Abad, F.J., & Garcia, L.F. (2000). Negligible sex differences in general intelligence. Intelligence, 28(1), 57-68.

Cromley, J. G. (2009). Reading achievement and science proficiency: International comparisons from the Programme on International Student Assessment. Reading Psychology, 30(2), 89-118. https://doi.org/10.1080/02702710802274903

de Zeeuw, E. L., de Geus, E. J. C., & Boomsma, D. I. (2015). Meta-analysis of twin studies highlights the importance of genetic variation in primary school educational achievement. Trends in Neuroscience and Education, 4(2015), 69–76. doi:10.1016/j.tine.2015.06.001.

Deary, I. J. (2012). Intelligence. Annual Review of Psychology, 63, 453-482.

Deary, I.J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. Intelligence, 35(1), 13-21.

Ding, H., & Homer, M. (2020). Interpreting mathematics performance in PISA: Taking account of reading performance. International Journal of Educational Research, 102, 101566. https://doi.org/10.1016/j.ijer.2020.101566

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What no child left behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. Journal of Educational Psychology, 104(2), 439-451. https://doi.org/10.1037/a0026280

Dueber, D. M., & Toland, M. D. (2021). A bifactor approach to subscore assessment. Psychological Methods. Advance online publication. https://doi.org/10.1037/met0000459

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. https://doi.org/10.3390/jintelligence6030042

Figazzolo, L. (2009). Impact of PISA 2006 on the Education Policy Debate, Education International. Available online at: http://pages.ei-ie.org/quadrennialreport/2009/s3.amazonaws.com/educationinternational/2009/assets/49/Impact_of_PISA_2006_EN.pdf

Fitzgerald, C. E., Estabrook, R., Martin, D. P., Brandmaier, A. M., & von Oertzen, T. (2021). Correcting the bias of the root mean squared error of approximation under missing data. Methodology, 17(3), 189-204.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. Psychological Science, 15(6), 373-378. https://doi.org/10.1111/j.0956-7976.2004.00687.x

Gagne, P., & Hancock, G. R. (2006). Measurement Model Quality, Sample Size, and Solution Propriety in Confirmatory Factor Models. Multivariate Behavioral Research, 41(1), 65–83. https://doi.org/10.1207/s15327906mbr4101_5

Grasby, K. L., Coventry, W. L., Byrne, B., Olson, R. K., & Medland, S. E. (2016). Genetic and environmental influences on literacy and numeracy performance in Australian school children in grades 3, 5, 7, and 9. Behavior Genetics, 46(5), 627–648. doi:10.1007/s10519-016-9797-z.

Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2016). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modelling of student achievement. The Annals of Applied Statistics, 10(4), 2405-2426. https://doi.org/10.1214/16-AOAS988.

Hart, S. A., Petrill, S. A., Thompson, L. A., & Plomin, R. (2009). The ABCs of math: A genetic analysis of mathematics and its links with reading ability and general cognitive ability. Journal of Educational Psychology, 101(2), 388-402. https://doi.org/10.1037/a0015115.

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.

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

Jez, S. J., & Wassmer, R. W. (2015). The impact of learning time on academic achievement. Education and Urban Society, 47(3), 284-306.

Kline, R. B. (2016). Principles and practice of structural equation modeling (Methodology in the social sciences). New York: The Guilford Press.

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. Intelligence, 36(2), 153-160. https://doi.org/10.1016/j.intell.2007.03.005.

Kovas, Y., Harlaar, N., Petrill, S. A., & Plomin, R. (2005). 'Generalist genes' and mathematics in 7-year-old twins. Intelligence, 33(5), 473-489. https://doi.org/10.1016/j.intell.2005.05.002

Kovas, Y., Haworth, C. M. A., Dale, P., Plomin, R., Weinberg, R. A., & Thomson, J. M. (2007). The genetic and environmental origins of learning abilities and disabilities in the early school years. Monographs of the Society for Research in Child Development, 72(3), i-156. https://doi.org/10.1111/j.1540-5834.2007.00439.x

Kovas, Y., Voronin, I., Kaydalov, A., Malykh, S. B., Dale, P. S., & Plomin, R. (2013). Literacy and numeracy are more heritable than intelligence in primary school. Psychological Science, 24(10), 2048-2056. doi:10.1177/0956797613486982.

Lemos, G. C., Abad, F. J., Almeida, L. S., & Colom, R. (2013). Sex differences on g and non-g intellectual performance reveal potential sources of STEM discrepancies. Intelligence, 41(1), 11-18.

Lynn, R., & Vanhanen, T. (2012). National iqs: A review of their educational, cognitive, economic, political, demographic, sociological, epidemiological, geographic and climatic correlates. Intelligence, 40(2), 226-234. http://dx.doi.org/10.1016/j.intell.2011.11.004

Marks, G. N. (2016). The relative effects of socio-economic, demographic, non-cognitive and cognitive influences on student achievement in Australia. Learning and Individual Differences, 49, 1-10. https://doi.org/10.1016/j.lindif.2016.05.012

Marks, G. N. (2021). Should value-added school effects models include student- and school-level covariates? Evidence from Australian population assessment data. British Educational Research Journal, 47(1), 181–204. doi:10.1002/berj.3684.

Marks, G. N., & O'Connell, M. (2021). Inadequacies in the SES–achievement model: Evidence from PISA and other studies. Review of Education, 9(3), e3293. https://doi.org/10.1002/rev3.3293.

Marks, G. N., & O'Connell, M. (2021). Inadequacies in the SES–achievement model: Evidence from PISA and other studies. Review of Education, 9(3), e3293. doi:https://doi.org/10.1002/rev3.3293.

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). Methods and Procedures: TIMSS 2019 Technical Report. International Association for the Evaluation of Educational Achievement.

O'Connell, M., & Marks, G. N. (2021). Are the effects of intelligence on student achievement and well-being largely

functions of family income and social class? Evidence from a longitudinal study of Irish adolescents. Intelligence, 84, 101511. https://doi.org/10.1016/j.intell.2020.101511

OECD & The World Bank (2015). The experience of middle-income countries participating in PISA 2000-2015. Paris, OECD Publishing.

OECD (2015). The ABC of gender equality in education: Aptitude, Behaviour, Confidence. Paris: OECD Publishing.

OECD (2019b) PISA 2018 technical report. Available online at: https://www.oecd.org/pisa/data/pisa2018technicalreport/ (Accessed 05 April 2022).

OECD (2019c), PISA 2018 Results (Volume I): What Students Know and Can Do, PISA, OECD Publishing, Paris, https://doi.org/10.1787/5f07c754-en.

OECD (2020). Learning time during and after school hours. in PISA 2018 Results (Volume V): Effective Policies, Successful Schools, OECD Publishing, Paris. https://doi.org/10.1787/639ec0b7-en

OECD. (2007). Science competencies for tomorrow's world. Volume 1. Paris: Organisation for Economic Co-operation and Development.

OECD (2012). PISA 2009 technical report. OECD Publishing.

OECD (2013). Technical report of the survey of adult skills (PIAAC). Paris: OECD.

OECD (2019). PISA 2018 Assessment and Analytical Framework. Paris: OECD Publishing.

Pagani, L., Fitzpatrick, C., Archambault, I., & Janosz, M. (2010). School readiness and later achievement: A French Canadian replication and extension. Developmental Psychology, 46, 984-994. Retrieved from https://doi.org/10.1037/a0018881.

Paige Harden, K. (2021). Genetic lottery. Why DNA matters for social equality. Princeton, Princeton University Press.

Perkins, D. N., & Grotzer, T. A. (1997). Teaching intelligence. American Psychologist, 52(10), 1125–1133. https://doi.org/10.1037/0003-066X.52.10.1125

Petrill, S. A. (2016). Behavioural genetic studies of reading and mathematics skills. In S. B. Malykh, Y. Kovas, & D. Gaysina (Eds.), Behavioural genetics for education (pp. 60-76). UK: Palgrave Macmillan.

Plomin, R., DeFries, J. D., Knopik, V. S., & Neiderhiser, J. M. (2013). Behavioral genetics (6th ed.). New York: Worth Publishers.

Pokropek, A., & Sikora, J. (2015). Heritability, family, school and academic achievement in adolescence. Social Science Research, 53(September), 73-88. doi:10.1016/j.ssresearch.2015.05.005.

Pokropek, A., Marks, G. N., & Borgonovi, F. (2021). How much do students' scores in PISA reflect general intelligence and how much do they reflect specific abilities? Journal of Educational Psychology. Advance online publication. https://doi.org/10.1037/edu0000687

Quinn, H. O. (2014). Bifactor Models, Explained Common Variance (ECV), and the Usefulness of Scores from Unidimensional Item Response Theory Analyses (Master Thesis). University of North Carolina, Chapel Hill. Retrieved from https://cdr.lib.unc.edu/concern/dissertations/w95051780

Raudonyte, I. (2019). Use of learning assessment data in education policy-making. IIEP-UNESCO Working Papers. Paris, International Institute for Educational Planning.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013b). Scoring and modeling psychological measures in the presence of multidimensionality. Journal of Personality Assessment, 95(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Bonifay, W., & Haviland, M. G. (2018). Bifactor Modelling and the Evaluation of Scale Scores. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development (pp. 675–707). Chichester: Wiley Blackwell. https://doi.org/10.1002/9781118489772.ch22

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. Journal of Personality Assessment, 92(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

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

https://doi.org/10.1177/0013164412449831

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

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, 21, 667-706.

Rindermann, H. (2008). Relevance of education and intelligence at the national level for the economic welfare of people. Intelligence, 36, 127–142. doi:10.1016/j.intell.2007.02.002.

Rindermann, H. (2018). Cognitive capitalism: Human capital and the wellbeing of nations. Cambridge: Cambridge University Press.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. American Sociological Review, 15, 351-357.

Robitzsch, A., Lüdtke, O., Goldhammer, F., Kroehne, U., & Köller, O. (2020). Reanalysis of the German PISA data: A comparison of different approaches for trend estimation with a particular emphasis on mode effects. Frontiers in psychology, 11, 884.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2015). Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. Journal of Personality Assessment, 98(3), 223–237. https://doi.org/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. https://doi.org/10.1037/met0000045

Rutkowski, D., Rutkowski, L., & Liaw, Y. L. (2018). Measuring widening proficiency differences in international assessments: Are current approaches enough? Educational Measurement: Issues and Practice, 37(4), 40-48.

Rutkowski, L., Rutkowski, D., & Liaw, Y. L. (2019). The existence and impact of floor effects for low-performing PISA participants. Assessment in Education: Principles, Policy & Practice, 26(6), 643-664.

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

Scheerens, J. and M. Hendriks (2014), “State of the Art of Time Effectiveness”, in Scheerens, J. (ed.), Effectiveness of Time Investments in Education: Insights from a review and meta-analysis, Springer International Publishing, Cham, http://dx.doi.org/10.1007/978-3-319-00924-7_2.

Schoon, E. Jones, H. Cheng, B. (2012). Maughan Family hardship, family instability, and cognitive development. Journal of Epidemiology and Community Health, 66, 716-722.

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.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. Intelligence, 35, 401-426

Wainwright, M. A., Wright, M. J., Luciano, M., Geffen, G. M., & Martin, N. G. (2005). Multivariate genetic analysis of academic skills of the Queensland core skills test and IQ highlight the importance of genetic g. Twin Research and Human Genetics, 8(6), 602-608. https://doi.org/10.1375/183242705774860259

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

Warne, R. T. (2020). In the know: Debunking 35 myths about human intelligence: Cambridge University Press.

Xia, Y., Yang, Y. (2019) RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. Behav Res 51, 409–428. https://doi.org/10.3758/s13428-018-1055-2

Xijuan Zhang & Victoria Savalei (2020) Examining the effect of missing data on RMSEA and CFI under normal theory full-information maximum likelihood, Structural Equation Modeling: A Multidisciplinary Journal, 27:2, 219-239, https://doi.org/10.1080/10705511.2019.1642111

Zaboski, B. A., II, 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. https://doi.org/10.1016/j.jsp.2018.10.001

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach’s α, Revelle’s β, and Mcdonald’s ωH: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70(1), 123–133.

https://doi.org/10.1007/s11336-003-0974-7

Bieber, T., & Martens, K. (2011). The OECD PISA study as a soft power in education? Lessons from Switzerland and the US. European Journal of Education, 46(1), 101-116. doi:10.1111/j.1465-3435.2010.01462.x.

Dobbins, M., & Martens, K. (2012). Towards an education approach à la finlandaise? French education policy after PISA. Journal of Education Policy, 27(1), 23-43. doi:10.1080/02680939.2011.622413.

Spearman, C. (1904). General intelligence objectively determined and measured. American Journal of Psychology, 15(2), 201–293. doi:10.2307/1412107.