Social Science Research 112 (2023) 102874

Contents lists available at ScienceDirect

Social Science Research

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

# From asking to observing. Behavioural measures of socio-emotional and motivational skills in large-scale assessments

F. Borgonovi [a,d,1,*], A. Ferrara [b], M. Piacentini [c]

[a] Social Research Institute, Institute of Education University College London, United Kingdom
[b] European University Institute, Italy
[c] Directorate for Education and Skills, Organisation for Economic Cooperation and Development, France
[d] OECD Centre for Skills, Organisation for Economic Cooperation and Development, France

ARTICLE INFO

*Keywords:*
Non-cognitive skills
Large-scale assessments
Behavioral measures
Comparability
PISA

ABSTRACT

Socio-emotional and motivational skills are routinely measured using self-reports in large-scale educational assessments. Measures exploiting test-takers' behaviour during the completion of questionnaires or cognitive tests are increasingly used as alternatives to self-reports in the economics of education literature. We compute behavioural measures of socio-emotional and motivational skills using data from the Programme for International Student Assessment (PISA). We find that these measures capture important aspects of students' academic profiles: some are importantly associated with contemporaneous performance and educational attainment and most measures have a high degree of stability over time. However, these measures are only limitedly correlated among themselves and have low correlations with self-report measures of the same constructs. This is likely a reflection of the fact that behavioural measures are representations of the test taker current 'state', rather than descriptions of the participant view of their own 'trait' like the self-report measures. Moreover, the low correlation across measures suggests that they capture different behavioural responses to the test-taking situation. These differences are still limitedly understood because the measures are constructed ex-post using collateral information collected during the administration of assessments rather than developed ex ante in line with theoretical models of human cognition and affect.

## 1. Introduction

Research in economics, psychology, sociology and education indicates that socio-emotional and motivational skills, also referred to as personality traits, temperament, non-cognitive skills or character skills in the literature, play an important role (and one independent from cognitive skills) in shaping individuals' long term outcomes (Duckworth and Seligman, 2005; Gutman and Schoon, 2016; Kautz et al., 2014; Roberts et al., 2007). Moreover, there is evidence that education policy and school-level practices have the potential to shape the acquisition of some of these skills (Heckman et al., 2013) although not all such skills may be amenable to the influence of external factors (Credé et al., 2017; Revelle, 2007).

* Corresponding author. Social Research Institute, University College London, 55Gordon Square London, WC1H 0NU United Kingdom.
*E-mail addresses:* f.borgonovi@ucl.ac.uk (F. Borgonovi), Alessandro.Ferrara@eui.eu (A. Ferrara), Mario.Piacentini@oecd.org (M. Piacentini).

[1] Francesca Borgonovi would like to acknowledge financial support from the British Academy through its Global Professorship scheme. The views expressed in the manuscript do not necessarily reflect those of the partner institutions. 4JS.

https://doi.org/10.1016/j.ssresearch.2023.102874
Received 19 April 2021; Received in revised form 24 December 2022; Accepted 4 March 2023
Available online 16 March 2023

The emphasis on socio-emotional and motivational skills - the term we adopt in this work - and the interest in how they can be promoted has prompted policy makers as well as researchers to consider how their measurement can best be integrated in benchmarking and accountability systems at international, regional and national levels. Standardised low-stakes large-scale tests such as the Programme for International Student Assessment (PISA), the Programme for International Assessment of Adult Competences (PIAAC), the IEA's Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS) as well as regional and national standardised tests have, in recent years, introduced the measurement of selected socio-emotional and motivational skills within questionnaires administered to participants, alongside the measurement of cognitive skills.

The increased attention devoted to understanding the role played by socio-emotional and motivational skills has also spurred a parallel and lively debate in the economics of education literature over what exactly low-stakes large-scale tests measure and the importance of factors such as test taking motivation as a determinant of differences in measures of cognitive skills across countries and population subgroups (Akyol et al., 2018; Balart and Oosterveen, 2019; Borghans and Schils, 2013; Borgonovi and Biecek, 2016; Brunello et al., 2018; Gneezy, et al., 2019; Wise and Kingsbury, 2002; Zamarro et al., 2019).

Questions remain on how best socio-emotional and motivational skills can be measured (Duckworth and Yeager, 2015; Heckman and Kautz, 2012). Likert-type scales administered to individuals through self-reported questionnaires are by far the most widely used instrument in large-scale settings. Self-reports have the advantage of being designed to reflect well-defined theoretical constructs and can be measured in a questionnaire with a relatively short time burden for respondents (Kyllonen and Kell, 2018). Moreover, self-reported measures of socio-emotional and motivational skills have been found to predict important education, work, and life outcomes (Soto, 2019, 2020; Wilmot and Ones, 2019). However, in applications of such assessments, practitioners often raise concerns about potential limitations of self-reports, such as misinterpretation, lack of information and memory bias, social desirability bias, response style bias and reference-group bias (Kankaraš, 2017) even if some researchers argue that these biases do not have a large influence on the association between personality measures and individual outcomes (Ones et al., 1996). Concerns over the measurement properties of self-reports raise questions as to whether the measurement of socio-emotional and motivational skills assessments could be complemented by additional measurement methods (Soland et al., 2019) since the use of multimethod assessments could provide more comprehensive information on the underlying constructs of interest (Meyer et al., 2001).

Laboratory and field experiments in which study participants are asked to complete tasks designed to reveal their socio-emotional and motivational skills in specific situations have been conducted at the international level (Cohn et al., 2019) but these instruments are expensive and burdensome. An alternative approach has been to use experiments to validate self-reports (see for example Falk and Hermle 2018).

Starting from research in psychometrics on test engagement (see Wise and Kingsbury, 2002 for a review), research in education and economics has developed measures of socio-emotional and motivational skills derived from observing and coding the behaviours of individuals when they participate in standardised educational assessment programmes (see Soland et al., 2019 for a review). Examples include item non-response rate for questionnaires (Zamarro et al., 2018) and, for computer-based tests, response time effort and rapid guessing (Goldhammer et al., 2014; Kuhfeld and Soland, 2020; Soland and Kuhfeld, 2019; Soland, 2019; Wise and Kong, 2005; Wise, 2017; Wise and Kingsbury, 2002). Behavioural measures share many characteristics with laboratory or field experiments and do not involve additional costs or burden for respondents but are theoretically ill-defined, because their construction is based on the behaviour of respondents while they perform a task that was designed with the intention of measuring a different set of constructs (Kroehne and Goldhammer, 2018). This is a typical challenge related to the use of behavioural data in social research (Salganik, 2019).

Our contribution is to connect research in psychology, educational measurement and economics by considering behavioural measures that have been used in the economics of education literature as indicators of four socio-emotional and motivational constructs – perseverance, self-regulation, endurance and conscientiousness (PSEC from now on) – and establish if they can complement established self-reported indicators of socio-emotional and motivational skills in large-scale assessment and monitoring systems. The use of task-based and performance-based measures has been promoted by researchers to avoid biases that may arise when relying in self-reports and because they do not entail additional burden and administration costs while conducting cross-country comparisons (Balart and Oosterveen, 2019; Borghans and Schils, 2013; Borgonovi and Biecek, 2016; Zamarro et al., 2019; Akyol et al., 2018). However, the valid use of behavioural measures as indicators of socio-emotional and motivational skills needs to be established because item features, context and framing might impede the use of test-takers' behaviours as indicators of broad psychological constructs. Since the validity of a measure for a particular purpose can never be absolutely proven, the typical validation approach is to develop a "validity argument" (Kane, 2006, 2013) that justifies the intended use by summarizing a compelling body of evidence from multiple sources. In the context of large-scale international educational assessments, the intended use of behavioural measures is to describe and monitor over time differences across countries and groups within countries in students' socio-emotional and motivational skills. We use data from the PISA study to establish a validity argument for such use based on theoretical expectations about relations to other variables. First, we evaluate to what extent the behavioural measures exhibit evidence of convergent validity (Fiske, 1971, p. 164): we expect behavioural measures to be significantly correlated with self-report measures that were developed to assess the PSEC construct. Second, we evaluate whether behavioural measures present similar differences across population sub-groups (i.e. gender, socio-economic status) as those observed through the use of self-reported measures, and examine the relationship of both groups of measures with performance in the PISA achievement test. For a subset of countries, we evaluate the evidence for predictive validity of the behavioural measures: the measures are expected to be correlated with future education outcomes, such as completion of higher education, that are influenced by socio-emotional skills (Danner et al., 2021). We also compare behavioural measures and self-report methods according to whether measures based on different survey rounds provide consistent conclusions, and conduct tests to identify how sensitive different instruments are to the use of different administration protocols or instruments.

Prior work in psychology has identified low correlations between self-reports and task-based performance measures as well as low

correlations between different task-based performance measures of socio-emotional constructs (see for examples Duckworth and Kern, 2011; Sharma et al., 2014; Cyders and Coskunpinar, 2011; Allom et al., 2016). However, no study to date has systematically examined the correlation between behavioural measures that are routinely used in the economics of education literature and that are developed using data from large-scale international assessments and self-reports routinely employed in these assessments.

Given the increasing academic and policy interest in examining disparities in socio-emotional and motivation skills across genders, students with a different socio-economic status (SES), as well as between students with and without an immigrant background (Greiff and Borgonovi, 2022; Fahle et al., 2019; OECD, 2021; Soland, 2018), we use PISA data to identify differences by gender, SES and immigrant background in behavioural and self-reported indicators of socio-emotional and motivational skills in different countries. Examining differences across population subgroups is important given persisting disparities in educationally relevant outcomes across genders, SES and immigrant background in both high (OECD, 2022) and low and middle income countries (Local Burden of Disease Educational Attainment Collaborators, 2020); the key role the provision of high quality education for all has in the Sustainable Development Goal framework, the role socio-emotional and motivational skills play in shaping educational outcomes (Almlund et al., 2011; Duckworth et al., 2007; Duckworth et al., 2012; Heckman et al., 2006; Poropat, 2009; Rosander and Bäckström, 2014) and the fact that different population groups may react differently to different measurement instruments (Soland, 2018; Borgonovi, 2021).

## 2. Data and methods

### 2.1. Data

We use data from PISA. PISA is a triennial large-scale international assessment of 15-year-old students conducted since 2000 and targeting the schooled population of children between the ages of 15 years and three months and 16 years and 2 months at the time of administration (15-year-olds from now on). It covers large, representative samples of 15-year-old students and over 80 countries have participated at least once since the first edition was conducted in 2000.[2] PISA is a low-stakes assessment at least at the individual level because no individual scores are released to students or schools. Therefore, intrinsic motivation plays a considerable role in guiding the behaviour of participants in the study. In the absence of external pressure and motivational drivers, variability in motivation is higher than in the presence of external incentives and the influence of PSEC on performance tends to be stronger (Barry et al., 2010; Cole et al., 2008; Wise and DeMars, 2010).

The core instruments of PISA are a 2-h long low-stakes assessment developed by international experts to test students' proficiency in reading, mathematics and science, and a 30-min background questionnaire. Greater information on PISA assessment instruments is available in the Online Annex A. We use data from the questionnaire to identify self-reported measures of PSEC and to derive three behavioural indicators and we use data from the assessment to derive two behavioural indicators. The PISA surveys are conducted on two-stage stratified representative samples of 15-year-old students enrolled in lower-secondary or upper secondary institutions [for details, see (Organisation for Economic Co-operation and Development OECD, 2009)]. We used sampling and replication weight in our analyses in line with recommended PISA procedures (OECD, 2014).

Most analyses are based on the PISA 2012 study. While more recent editions of PISA are available, 2012 is the only edition that contains design features that allow to validate alternative operational definitions of the behavioural measures and test the robustness of behavioural measures to different administration conditions. We use PISA 2003 data to examine the stability of measures over time and PISA 2000 for Canada, Denmark and Switzerland and PISA 2003 for Australia to examine how predictive measures are of educational attainment at age 25. In these countries, PISA participants were followed in national longitudinal studies. The follow-up data for Denmark is derived from PISA and the OECD Survey of Adult Skills (PIAAC). For Switzerland, we used data from the Transitions from Education to Employment (TREE1) surveys. For Australia, we used data from the Longitudinal Study of Australian Youth (LSAY). For Canada, we used data from the Youth in Transition Survey (YITS). Canadian data have not been released for public access, so all estimates were obtained through collaborations with national researchers at Statistics Canada, who conducted the analyses on the basis of statistical programmes prepared by the authors. While the frequency and timeline of follow-up surveys vary by country, all the datasets collect information on young people when they were approximately 25 years old (the only exception is Denmark, where the analysis refers to individuals who were either 26 or 27 years old). Table B1 in the Supplementary Online Annex provides additional details on the longitudinal studies.

### 2.2. PSEC measures

#### 2.2.1. Self-reported measures

*2.2.1.1. The PISA index of lack of perseverance.* The first measure of noncognitive skills based on students' self-reports is the index of perseverance. The index is derived using five questions from the core PISA student questionnaire that ask participants to report the

[2] PISA 2012had a total of 64 participating countries: Albania, Argentina, Australia, Austria, Belgium, Brazil, Bulgaria, Canada, Chile, Chinese Taipei, Colombia, Costa Rica, Croatia, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hong Kong, Hungary, Iceland, Indonesia, Ireland, Israel, Italy, Japan, Jordan, Kazakhstan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Macao (China), Malaysia, Mexico, Montenegro, Netherlands, New Zealand, Norway, Peru, Poland, Portugal, Qatar, Romania, Russia, Serbia, Shanghai, Singapore, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Thailand, Tunisia, Turkey, United Arab Emirates, United Kingdom, United States, Uruguay and Vietnam.

extent to which they felt that the following statements described them: "When confronted with a problem, I give up easily"; "I put off difficult problems"; "I remain interested in the tasks that I start"; "I continue working on tasks until everything is perfect"; "When confronted with a problem, I do more than is expected of me". Possible answers were "Very much like me", "Mostly like me", "Somewhat like me", "Not much like me" and "Not at all like me". The index was recoded so that higher values indicate lower student perseverance. The median Chronbach alpha across OECD countries was 0.8.

*2.2.1.2. Measures from the PISA effort thermometer.* The second self-reported measure that we consider was derived from the PISA effort thermometer. In the last page of the PISA assessment booklet, students who sat the assessment were asked how much effort they put in the test and to consider how much effort they would have put in the test if their performance in PISA had counted towards their school marks. Responses could range from 1 to 10 with 10 being maximum effort. We include the answers to the question on the effort students reported having put in the PISA test and recoded it so that a higher value signals lower effort. We also consider a second index, which is the difference between the effort that would have been invested if the test were marked and the actual effort invested. Students with a greater score on such index should be seen as being less conscientious or intrinsically motivated, since they reported having invested less effort in the PISA test compared to how much effort they would have invested had the PISA test been associated with an extrinsic reward, i.e. a school grade. Although both indicators are categorical, they were treated as continuous in analyses since they have a comparatively large number of categories and treating them as continuous allows for easier estimation and interpretation.

### 2.2.2. Questionnaire-based behavioural indicators

We consider the following questionnaire-based behavioural indicators based on the core student questionnaires[3]: item non-response, non-differentiation and inconsistency.

*2.2.2.1. Item non-response rate.* Survey item non-response is considered to provide information on the level of effort and motivation of students (Hitt et al., 2016). The indicator is constructed using multiple-choice questions for which an answer was possible and applicable to the responding student (we dropped items where the response is conditional on prior items). The item non-response rate indicator is defined as:

$$Item\ non-response = \frac{Number\ of\ applicable\ items\ left\ blank}{Number\ of\ applicable\ items}$$

Higher values indicate that students left more questions unanswered.

*2.2.2.2. Inconsistency.* The inconsistency indicator (also called response non-consistency in the literature) is a measure of careless answering increasingly used in the literature (Hitt, 2015; Wise and DeMars, 2006; Zamarro et al., 2019). The measure is based on item-rest correlations and it relies on the assumption that, in an internally consistent scale, the answer to a specific item should be correlated with the answers to the rest of the items that make up that scale. A student who answers a specific item in a way that is unpredictable based on the answers to other items in the same scale is thus considered to provide a "careless answer". The response non-consistency indicator is constructed using Likert-type items belonging to a scale meant to measure some latent attitude or behaviour of respondents. First, we adjusted answers of reverse-coded items (such as negatively-phrased items in positively-phrased scales) and then for every item, we regressed the item response on the average of answers in the remaining items belonging to the same scale. We run regressions separately for each country in our sample to account for the fact that the internal consistency of scales might differ across countries and considered only respondents who had non-missing responses for at least 3 items in a scale. For each item, we fit the following linear model:

$$Y_{ijs} = \beta_0 + \beta_1 \overline{X}_{is,-j} + \eta_{ijs}$$

Where $Y_{ijs}$ represents the answer of student $i$ to item $s$ in scale $j$ and $\overline{X}_{is,-j}$ represents the average response on all remaining items in the scale, $\beta_0$ represents a constant and $\eta_{ijs}$ is an item-specific, scale-specific and student-specific error term representing the degree to which student $i$ gave an unpredictable answer to item $s$ in scale $j$. These sets of regressions are equivalent to item-rest correlations common in psychometrics (Hitt, 2015). For each student, we averaged the absolute values of all residuals to obtain a measure of non-consistency in their answers across the entire PISA survey. Higher values imply that a student gave more inconsistent answers than another student. In PISA 2012, we examine 93 items across 17 scales in the main questionnaire.

*2.2.2.3. Non-differentiation.* The non-differentiation indicator reflects the extent to which students tend to select the same response across a set of similar and related items. This can be the result of a careful analysis of questions, but also of satisficing behaviour. Respondents could realise that the items in a certain set are similar and, in order to minimise the effort exerted, give the same response to all. Several studies have found support for this claim: non-differentiation is more common among less educated individuals and

[3] The average mean across countries is not precisely 0 because the z score standardization was conducted on the weighted pooled sample while the mean across countries represents the arithmetic mean of weighted country specific samples,i.e. while the country specific mean reflects the country specific population distribution, countries with large overall populations contribute to the same extent to the average as do countries with small populations. By contrast, in the pooled sample, these countries contribute more.

towards the end of a questionnaire compared to the beginning (Knowles, 1988; Krosnick, 1991; Vannette and Krosnick, 2014). We focus only on Likert-type item sets in PISA, since for this type of questions the likelihood of mindful respondents not differentiating their answers is lower. We limit the analysis to item sets having at least 4 items (a total of 16 in PISA 2012) and to students having non-missing responses for at least 3 items. In the most stringent version of our measure, a student is considered to be non-differentiating within an item set if he or she gave the same answer to all items in the item set. To compute a non-differentiation metric for each student, we compute the percentage of item sets in which they did not differentiate their answers out of all valid items sets.

$$Non-differentiation = \frac{Number\ of\ applicable\ items\ sets\ without\ differentiation}{Number\ of\ applicable\ item\ sets}$$

Since absolute non-differentiation can be too stringent, we decided to also adopt a more lenient version of the index (Barge and Gehlbach, 2012). We relaxed the definition such that a student is considered not to differentiate within an item set when he or she gives the same response in all but one item in the set. All analyses were replicated using this index and are available from the authors upon request.

### 2.2.3. Test-based behavioural indicators

*2.2.3.1. Decline in performance.* The core cognitive test in PISA 2012 was delivered as a paper and pencil test. This was designed to last 2 h in total for each participant and was organised around a series of clusters of test questions. Each cluster was designed to take about 30 min to complete. Each student was randomly allocated a booklet containing four clusters of test questions and a total of 13 booklets were administered in 2012. Each booklet contained different clusters, which were rotated across the booklets so that each cluster was administered at least once with any other cluster and each cluster appeared at least once in one of the four potential positions within the booklet. Table C1 in the Supplementary Annex illustrates the design of the standard paper and pencil booklets in PISA 2012.

On average, in various PISA waves, the number of correct responses students give declines across the test: students tend to display higher performance in the first cluster than in the second and so on (Borghans and Schils, 2013; Zamarro et al., 2019). While different strategies have been used to develop PSEC indicators based on decline in performance in low-stakes standardised assessments (Borghans and Schils, 2013; Borgonovi and Biecek, 2016; Brunello et al., 2018; Zamarro et al., 2019) we develop an individual level indicator following Zamarro et al. (2019). The assumption behind the measure is that students with high levels of socio-emotional and motivational skills are more likely to maintain a similar level of performance throughout the test while those with lower levels of socio-emotional and motivational skills are more likely to display steep declines in performance as a function of item position. We only consider performance in the first three clusters to deal with end of test non-response. Some students in fact fail to reach items at the end of the test, which could lead to biased estimates (Author, 2016; Debeer and Janssen, 2013).

To obtain our measure, we estimate the following linear random coefficient model:

$$y_{ij} = \delta_0 + \delta_0^i + \delta_1 Q_{ij} + \delta_1^i Q_{ij} + \gamma_j + \theta_j + \varepsilon_{ij}$$

Where $y_{ij}$ is equals zero if respondent $i$ answered incorrectly to question $j$ and 1 if he or she answered correctly or if he or she got partial credit for his/her answers.[4] $Q_{ij}$ represents the position of question $j$ rescaled for each student such that the first question takes value zero and the last item in the third cluster takes value 1. $\delta_0$ represents the average student's performance on the first question in the test and $\delta_1$ is the average performance drop from the first question to the last. $\gamma_j$ are question fixed effects, which allow us to control for question difficulty and nature (such as multiple choice versus open-ended question). $\theta_j$ are booklet fixed effects to control for the sequence of clusters in the booklet (for example starting with a math or reading cluster). $\delta_0^i$ is a random intercept and $\delta_1^i$ is a random coefficient that allow for students to deviate from the average values.

The model was estimated separately for each country using Maximum Likelihood methods and allowing maximum flexibility in the covariance matrix for random effects (all variances and covariances could be distinctly estimated). Standard errors were clustered at the school level, to account for the clustered nature of the PISA samples. Fitting the model produced estimates of the standard deviation of random effects ($\delta_0^i$ and $\delta_1^i$) and the correlation between them. We used these to predict the best linear unbiased predictions (BLUPS) for the random effects.

The index of performance decline corresponds to BLUP estimate of the random slope parameter $\delta_1^i$, which measures individuals' decline in performance between the first and the last question (in the third cluster), accounting for question difficulty and the booklet that the student was assigned.

The PISA 2000 test design was different from subsequent editions in that the major domain (reading) was assessed significantly more in depth than the other two. Seven booklets began with three reading clusters and ended with a different domain. To maintain a

[4] Each PISA round includes a core student questionnaire, as well as a series of optional questionnaires designed to gather for a restricted number of interested countries, additional information. We construct two sets of questionnaire-based behavioural indicators. The first is based solely on responses to the core questionnaires so that we have valid measures for all PISA-participating countries in each round. We present these results in the manscript. In addition, we developed all indicators using both the core questionnaire as well as the optional Information and Computer Technology (ICT) questionnaires in order to maximise the observations used to construct the indices for each individual (for the restricted set of countries that administered also the optional questionnaires). Results are in line with those presented and are available from the authors upon request.

balanced sample, we decided to focus our analyses only on those booklets. Therefore, our measure in PISA 2000 is entirely based on performance decline in the reading assessment. Higher values in the indicator indicate that students had a steeper performance decline.

*2.2.3.2. Response time effort.* In 2012 some countries administered an optional computer-based assessment on top of the paper-based assessment instruments, generating data on participants' interactions with the testing platform, including a timestamp to mark each interaction. We follow Wise and Kong (2005) and develop a response-time effort indicator using the log files for the PISA 2012 computer-based assessment. The indicator constitutes the proportion of items, out of the total number of items in the test, on which respondents spent less than a threshold time T. We computed the measure using thresholds of 3, 5, and 10 s as well as setting the threshold at 10% of the mean item duration with a maximum threshold of 10 s as proposed by Wise and Ma (2012), and our results were robust across these different specifications. For the sake of conciseness, we report most of our results only for the measure using the middle threshold of 5 s (results for other thresholds are available upon request). Higher values in the indicator signal that students exerted less effort in their responses.

### *2.2.4. Robustness of the measures with respect to test design*

We performed a number of robustness checks to assess the questionnaire-based and test-based behavioural indicators, exploring how they might vary according to certain features of the PISA assessment.

In 2012 the background questionnaire had a rotation design and therefore different students were randomly allocated different questionnaire materials (scales with a different number of items or with a different length of the prompts or the same material but placed in different positions within the questionnaire). In Supplementary Annex D we illustrate the rotation design and provide estimates of how much the value of different questionnaire-based behavioural indicators depends on the questionnaire set students were administered. We find that non-differentiation and inconsistency appear to be sensitive while non-response is less dependent on the specific questions contained in the questionnaire. Given the structure of our data, we could separately assess the robustness of indicators to specific features of the questionnaires such as their length or the type and order of items they contained.

We performed other robustness checks for the test-based indicator of performance decline. We examined variations in the performance decline indicator depending on the content of the test (amount of reading and mathematics material and position of mathematics and reading items). Results in Supplementary Online Annex C suggest that the performance decline is not related to the structure of the assessment booklet. In order to identify the robustness of the indicator to mode of administration, in Online Annex E we estimated differences in performance decline across the PISA 2012 computer-based assessment and the PISA 2012 paper-based assessment. Results indicate that students' ranking in the performance decline measure based on the computer-based test is positively related to their ranking in paper-based performance decline.

### *2.2.5. Indicator standardization and distributions*

In order to draw meaningful comparisons across indicators in our analyses, we recoded each indicator such that a higher value indicates a lower level of PSEC. We also standardised each indicator such that it is set to have a mean of zero and a standard deviation of one across the pooled student population. Figure F1 in the Supplementary Online Annex presents the distribution of the PSEC measures. It reveals that the item non-response indicator and the response time effort indicator have highly right-skewed distributions. This is also the case, to a less extent, for the inconsistency indicator. In order to address this, we ran all analyses using the original scale as well as the log-transformed versions of these variables for robustness. Since results using the two sets of measures were aligned we decided to present results in the original metric in the manuscript because these are more easily interpretable. Robustness results are available upon request from the authors.

## *2.3. Background variables*

In a set of analyses, we analyse differences in PSEC indicators by students' gender, socio-economic background and immigrant background. These variables were constructed using information gathered in the PISA student questionnaire. Socio-economic background (SES) is defined using the PISA index of economic, social and cultural status The index is an aggregate indicator reflecting differences across students in parental educational attainment, parental occupational status and household resources (see Organisation for Economic Co-operation and Development (OECD), 2014) for a detailed description of the index). We provide results comparing students in the top and bottom quartile of their national distribution. Students are defined as having an immigrant background if their parents were born outside the country in which they sat the PISA test; in the analysis, these students are compared with those with two native-born parents.

In addition to these variables, in the longitudinal analyses, we also control for other variables gathered in the PISA background questionnaire: students' school grade and age when they sat the PISA assessment, and whether respondents reported most frequently speaking the language of the PISA assessment at home. We also control for their achievement in the first cluster of the assessment (using an indicator of the average percentage of correct responses in the first booklet).

## *2.4. Methods*

In the first set of analyses, we investigate the relationship between PSEC indicators at the individual- and country-levels, as well as

their country-level stability over time. We provide descriptive statistics for each PSEC indicator at the country-level and compare countries' averages and variations for each indicator. We then measure the correlation between all behavioural and self-reported PSEC measures using both Pearson and Spearman correlations. We run analyses at the individual level within each country and also at the country level. For robustness, we log-transformed count variables that were right-skewed (non-response rate, response time effort and inconsistency). Next, we investigate the stability of the indicators over time, by comparing country scores in each PSEC indica tor in PISA 2003 and2012, the only two waves including at least one self-reported instrument, the effort thermometer. We measure the Pearson and Spearman correlations between countries' average scores, as well as their relative ranking, for each PSEC indicator across the two waves.

In a second set of analyses we investigate country-level differences in PSEC indicators across students' background characteristics, including gender, socio-economic status and immigrant background. We computed T-tests and report the statistical significance of group differences.

In a third set of analyses, we evaluate how strongly PSEC indicators are related to contemporaneous achievement by fitting the following models:

$$y_{ik} = \delta_0 + \delta_1 Q_{ik} + \gamma_k + \varepsilon_{ij} \quad \text{(A)}$$

$$y_{ik} = \delta_0 + \delta_1 Q_{ik} + \delta_2 V_{ik} + \gamma_k + \varepsilon_{ij} \quad \text{(B)}$$

$$y_{ik} = \delta_0 + \delta_1 Q_{ik} + \delta_2 V_{ik} + \delta_3 U_{ik} + \gamma_k + \varepsilon_{ij} \quad \text{(I)}$$

Baseline models described by equation (A) estimate the association $\delta_1$ between each achievement domain $y$ (reading, mathematics, science and problem solving) fortudentt $i$ in country $k$ and each PSEC indicator $Q$ without controls (except for country fixed effects $\gamma$). Models described by equation (B) estimate the association between each achievement domain and each PSEC indicator controlling for country fixed effects as well as a set of additional control variables $V$ including grade level, age, gender, immigrant background, socio-economic status, and if the student was a native-speaker of the language in which the PISA assessment was delivered. For behavioural measures, Models described by equation (C) further control for the self-reported perseverance measure $U$, to assess whether the relationship between behavioural indicators of PSEC and performance is robust to inclusion of self-reported measures of PSEC. All models include students with non-missing values for all PSEC indicators and control variables to ensure comparability of results across specifications.

Finally, for the subset of countries with available longitudinal data, we analyse the predictive validity of PSEC indicators. We measure the associations between the indicators measured in PISA at the age of 15 and the likelihood of completing upper secondary education (academic or vocational) and of completing a university degree by the age of 25. We estimate the following two linear probability models:

$$y_{it} = \delta_0 + \delta_1 Q_{it-10} + \delta_2 V_{it-10} + \varepsilon_{ijt} \quad (1)$$

$$y_{it} = \delta_0 + \delta_1 \overline{Q}_{it-1} + \delta_2 V_{it-10} + \delta_3 u_{it-10} + \varepsilon_{ijt} \quad (2)$$

In model (1) we regress the two binary outcome variables $y$ (having completed high-school and having completed university by age 25) on one PSEC indicator ($Q$) at a time, and a set of control variables $V$ including students' school grade, age, gender, immigrant background, socio-economic status,[5] whether respondents were native-speakers at the time of the PISA assessment as well as for their achievement in the first cluster of the assessment (using an indicator of the average percentage of correct responses in the first booklet). The latter was taken as a proxy of their cognitive skills and was preferred to actual assessment scores because of the possibility of collinearity with the index of performance decline. In model (2), we estimate the relationship between the education completion at 25 and the set of behavioural indicators of PSEC $\overline{Q}$ (performance decline, inconsistency, non-response, non-differentiation), when accounting for self-reported measures of PSEC. For Denmark and Switzerland, model (2) includes a self-reported index perseverance (u), obtained from students responses to questions that were administered as part of an optional Cross-Curricular Competences questionnaire (CCC) in PISA 2000 (see Supple mentary Online Annex M). For Australia, the only self-reported measure of PSEC available is the effort thermometer, so model 2 for Australia includes this measure as additional control. No self-reported measure of PSEC is available for Canada. Results are similar when we fit logistic regression models and consider odds ratios. However, since we are interested in comparing estimates across different models and across different countries, we decided to report estimates obtained using linear probability since it has been suggested that odds ratios cannot be compared meaningfully across samples and models (Mood, 2010).

[5] In some test items respondents could obtain a partial credit, for example when the final response provided was incorrect (because of a typo or small calculation mistake) but the respondent correctly followed the procedure to solve an item. For simplicity and in line with most research in this area we coded these answers as 1 (correct).

**Table 1a**
Individual level Pearson correlations between PSEC indicators.

| | Self-reported lack of perseverance | Self-reported lack of effort in test | Difference in effort (PISA-marked) | Non-response | Non-differentiation | Non-differentiation (more lenient) | Inconsistency | Performance decline | Response time effort (3 s) | Response time effort (5 s) | Response time effort (10 s) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Self-reported lack of perseverance | 1.00 | | | | | | | | | | |
| Self-reported lack of effort in test | 0.21*** | 1.00 | | | | | | | | | |
| Difference in effort (PISA-marked) | 0.10*** | 0.69*** | 1.00 | | | | | | | | |
| Non-response | 0.07*** | 0.08*** | 0.02*** | 1.00 | | | | | | | |
| Non-differentiation | −0.07*** | −0.01* | −0.05*** | 0.03*** | 1.00 | | | | | | |
| Non-differentiation (more lenient) | −0.07*** | −0.01*** | −0.05*** | 0.01*** | 1.00*** | 1.00 | | | | | |
| Inconsistency | 0.14*** | 0.10*** | 0.06*** | 0.12*** | −0.33*** | −0.33*** | 1.00 | | | | |
| Performance decline | 0.05*** | 0.07*** | 0.05*** | 0.01*** | 0.05*** | 0.05*** | −0.00 | 1.00 | | | |
| Response time effort (3 s) | 0.01* | 0.05*** | 0.02*** | 0.04*** | 0.04*** | 0.04*** | 0.03*** | 0.03*** | 1.00 | | |
| Response time effort (5 s) | 0.03*** | 0.06*** | 0.04*** | 0.06*** | 0.09*** | 0.08*** | 0.05*** | 0.05*** | 0.74*** | 1.00 | |
| Response time effort (10 s) | 0.05*** | 0.07*** | 0.03*** | 0.10*** | 0.14*** | 0.13*** | 0.09*** | 0.08*** | 0.45*** | 0.74*** | 1.00 |

Note: PISA 2012data. The number of individual level observations is 224 748 for results involving indicators obtained from the paper-based assessment and 59 190 for results involving indicators obtained from the computer-based assessment. All correlations are statistically significant at the 0.001 level, except for the one between response time effort (3 s) and self-reported lack of perseverance, and between self-reported lack of effort in test and non-differentiation at the 0.10 level.

**Table 1b**
Individual level Spearman correlations between PSEC indicators.

| | Self-reported lack of perseverance | Self-reported lack of effort in test | Difference in effort (PISA-marked) | Non-response | Non-differentiation | Non-differentiation (more lenient) | Inconsistency | Performance decline | Response time effort (3 s) | Response time effort (5 s) | Response time effort (10 s) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Self-reported lack of perseverance | 1.00 | | | | | | | | | | |
| Self-reported lack of effort in test | 0.23 | 1.00 | | | | | | | | | |
| Difference in effort (PISA-marked) | 0.13 | 0.72 | 1.00 | | | | | | | | |
| Non-response | 0.11 | 0.1 | 0.01 | 1.00 | | | | | | | |
| Non-differentiation | −0.06 | −0.03 | −0.06 | −0.02 | 1.00 | | | | | | |
| Non-differentiation (more lenient) | −0.06 | −0.03 | −0.06 | −0.02 | 1 | 1.00 | | | | | |
| Inconsistency | 0.11 | 0.08 | 0.06 | 0.12 | −0.4 | −0.4 | 1.00 | | | | |
| Performance decline | 0.06 | 0.07 | 0.04 | 0.05 | 0.04 | 0.04 | 0.01 | 1.00 | | | |
| Response time effort (3 s) | 0.02 | 0.05 | 0.03 | 0.04 | 0 | 0.02 | 0.04 | 0.04 | 1.00 | | |
| Response time effort (5 s) | 0.03 | 0.05 | 0.02 | 0.06 | 0.05 | 0.05 | 0.07 | 0.06 | 0.54 | 1.00 | |
| Response time effort (10 s) | 0.06 | 0.08 | 0.05 | 0.09 | 0.08 | 0.08 | 0.1 | 0.1 | 0.29 | 0.53 | 1.00 |

Note: PISA 2012data. The number of individual level observations is 224 748 for results involving indicators obtained from the paper-based assessment and 59 190 for results involving indicators obtained from the computer-based assessment. All correlations are statistically significant at the 0.001 level.

**Table 2**
Correlations between country level PSEC indicators measured in PISA 2003and PISA 2012

| | Absolute correlation | Rank correlation |
|---|---|---|
| Self-reported lack of effort in test | 0.842 | 0.755 |
| Difference in effort (PISA-marked) | 0.86 | 0.704 |
| Non-response | 0.282 | 0.525 |
| Non-differentiation | 0.845 | 0.848 |
| Inconsistency | 0.897 | 0.88 |
| Performance decline | 0.405 | 0.526 |
| Instrumental motivation | 0.945 | 0.945 |
| Sense of belonging at school | 0.714 | 0.724 |

Note: PISA 2012and 2003 data. Absolute correlations for Non-response and Performance decline were computed after removing two countries that were outliers in 2012 (Norway for non-response and Brazil for Performance decline). Figures G1 to G12 in the Supplementary Online Annex illustrate the correlations between indicators measured in 2003 and in 2012.

## 3. Results

### 3.1. Individual and country level correlations of PSEC indicators

We provide country-level descriptive statistics for PSEC indicators in Supplementary Annex G Table G1 illustrates country-specific results for means, Table G2 illustrates country specific results for standard deviations and Table G3 reports country level correlations between different measures. In order to illustrate differences across countries in average levels of PSEC indicators we compare, for each indicator, if the country specific mean was in line, above or below the estimated mean across all the countries that took part in the analysis.[6] Results indicate that at the country level the correlation between different PSEC indicators is weak. Countries with comparatively high mean levels on some indicators have comparatively low levels on other indicators.

Table 1a and 1b indicate that individual level associations between different PSEC measures are generally weak although almost all associations are in the expected directions. Correlations are higher between the three self-reported measures: Pearson's $r = 0.21$ and Spearman's $\rho = 0.23$ between lack of perseverance and lack of effort in the PISA test; $r = 0.69$ and $\rho = 0.72$ between lack of effort in the PISA test and the difference between this measure and this measure in a hypothetical PISA test that counted towards their grade; and r $= 0.10$ and $\rho = 0.13$ between lack of perseverance and the difference in effort between the PISA low-stake condition and the hypothetical higher stake condition.

Meta-analytic reviews of cross-method convergence of several psychological constructs have found weak correlations between informant-report and task measures (Meyer et al., 2001). Our estimates are comparable to some of Meyer's result..

In Table 3 we present country-level Pearson correlations between each PSEC indicator measured in the PISA waves of 2012 and in 2003 (for scatterplots see Supplementary Annex H) (see Table 2). We compute correlations for countries' average scores and their ranking in each indicator. We also present correlations for the PISA indexes of sense of belonging and instrumental motivation, which we used as a benchmark of changes over time in self-reported constructs in a non-PSEC measure. Although successive cohorts of students might have different levels of PSEC, we do not expect large differences over a nine-year period. Absolute correlation coefficients are all above 0.7 except for the non-reponse indicator ($r = 0.452$) and highlight a high degree of stability for measures that are based on different questionnaire and assessment instruments but that are aligned in terms of administration conditions and underlying framework.

### 3.2. Differences across student background characteristics

In Fig. 1 we present gender differences in the PSEC indicators for the average of the countries in our sample. Positive values indicate that females have higher z-values than males (and thus worse PSEC outcomes), while negative values indicate that males have higher z-values than females (and thus worse PSEC outcomes). Fig. 2 shows results by socio-economic status (positive values indicate worse PSEC outcomes for socio-economically advantaged students), and Fig. 3 results by immigrant background (positive values indicate worse PSEC outcomes for students without an immigrant background). Results on differences across gender, socio-economic background and immigrant background for the full set of countries with available data can be found in Tables I2, J2 and, K2 in the Supplementary Online Annex.

On average, the gender gap in the standard self-reported lack of perseverance, although statistically significant, is quantitatively very small: 15-year-old female students report slightly more lack of perseverance than 15-year boys ($d = 0.03$). By contrast, on all the behavioural indicators, females displayed lower overall z scores than males (i.e. had better PSEC outcomes). In the questionnaire, they were less likely than males to provide inconsistent response patters ($d = -0.11$), to fail to provide answers ($d = -0.11$) and to provide

[6] In PISA 2000, the PISA index of Economic Social and Cultural Status (ESCS) had not yet been developed, so instead, we use three indicators that were available at the time – the one of parental highest educational attainment, highest occupational status and cultural possessions at home. These three are almost identical to the three indicators that are used to construct the ESCS index (only the indicator of home possessions differs slightly).

**Table 3**
The relationship between PSEC indicators and achievement.

| Indicator | Model | Reading | | | Math | | | Science | | | Collaborative problem-solving | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | b | sig | Adjusted R-squared | b | sig | Adjusted R-squared | b | sig | Adjusted R-squared | b | sig | Adjusted R-squared |
| Self-reported lack of perseverance | A | −0.14 | *** | 0.0069 | −0.17 | *** | 0.007 | −0.16 | *** | 0.0056 | −0.14 | *** | .0075 |
| | B | −0.11 | *** | 0.3 | −0.13 | *** | 0.31 | −0.11 | *** | 0.3 | −0.10 | *** | .19 |
| Self-reported lack of effort in test | A | −0.13 | *** | 0.0000024 | −0.10 | *** | 0.0014 | −0.11 | *** | 0.00072 | −0.09 | *** | 0.000015 |
| | B | −0.11 | *** | 0.26 | −0.10 | *** | 0.26 | −0.10 | *** | 0.26 | −0.10 | *** | .18 |
| Difference in effort (PISA-marked) | A | −0.02 | *** | 0.0083 | −0.02 | *** | 0.0098 | −0.02 | *** | 0.0096 | −0.03 | *** | .00095 |
| | B | −0.04 | *** | 0.26 | −0.05 | *** | 0.26 | −0.05 | *** | 0.25 | −0.06 | *** | .17 |
| Non-response | A | −0.37 | *** | 0.07 | −0.33 | *** | 0.062 | −0.34 | *** | 0.066 | −0.34 | *** | .049 |
| | B | −0.27 | *** | 0.33 | −0.26 | *** | 0.33 | −0.27 | *** | 0.33 | −0.27 | *** | .21 |
| | C | −0.32 | *** | 0.33 | −0.32 | *** | 0.33 | −0.32 | *** | 0.33 | −0.33 | *** | .22 |
| Non-differentiation | A | −0.14 | *** | 0.036 | −0.09 | *** | 0.023 | −0.11 | *** | 0.029 | −0.12 | *** | .011 |
| | B | −0.10 | *** | 0.32 | −0.08 | *** | 0.31 | −0.09 | *** | 0.31 | −0.10 | *** | .2 |
| | C | −0.13 | *** | 0.32 | −0.09 | *** | 0.32 | −0.11 | *** | 0.32 | −0.12 | *** | .2 |
| Inconsistency | A | −0.14 | *** | 0.014 | −0.13 | *** | 0.013 | −0.13 | *** | 0.011 | −0.13 | *** | .024 |
| | B | −0.09 | *** | 0.31 | −0.10 | *** | 0.31 | −0.09 | *** | 0.31 | −0.10 | *** | .2 |
| | C | −0.08 | *** | 0.31 | −0.09 | *** | 0.31 | −0.07 | *** | 0.31 | −0.09 | *** | .2 |
| Performance decline | A | −0.15 | *** | 0.0027 | −0.14 | *** | 0.00093 | −0.14 | *** | 0.0016 | −0.18 | *** | .0076 |
| | B | −0.13 | *** | 0.31 | −0.12 | *** | 0.31 | −0.12 | *** | 0.3 | −0.15 | *** | .2 |
| | C | −0.12 | *** | 0.31 | −0.12 | *** | 0.31 | −0.12 | *** | 0.31 | −0.14 | *** | .2 |
| Response time effort (5 s) | A | −0.19 | *** | 0.021 | −0.16 | *** | 0.016 | −0.18 | *** | 0.017 | −0.23 | *** | .032 |
| | B | −0.15 | *** | 0.21 | −0.13 | *** | 0.19 | −0.15 | *** | 0.19 | −0.20 | *** | .18 |
| | C | −0.13 | *** | 0.2 | −0.12 | *** | 0.19 | −0.14 | *** | 0.2 | −0.19 | *** | .18 |

Notes: The number of observations in models A and B testing the significance of self-reported indicators for math, science and reading is: 247 598 (self-reported lack of perseverance) 340 096 (self-reported lack of effort) and 340 096 (difference in effort (PISA-marked). For collaborative problem-solving it is: 150 252 (self-reported lack of perseverance) 210 212 (self-reported lack of effort) and 210 212 (difference in effort (PISA-marked)).

Models B and C control for students' school grade, age, gender, socio-economic status, immigrant background and linguistic background. For indicators obtained from the paper-based assessment, the number of observations is 375 657 in models A and B and 247 598 in model C for math, science and reading (227 316 in models A and B and 150 252 in model C for collaborative problem-solving). For the indicator obtained from the computer-based assessment, the number of observations is 48 502 in A and B and 31 982 in model C for math, science and reading (48 502 in A and B and 31 982 in model C).
sig= significance level *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

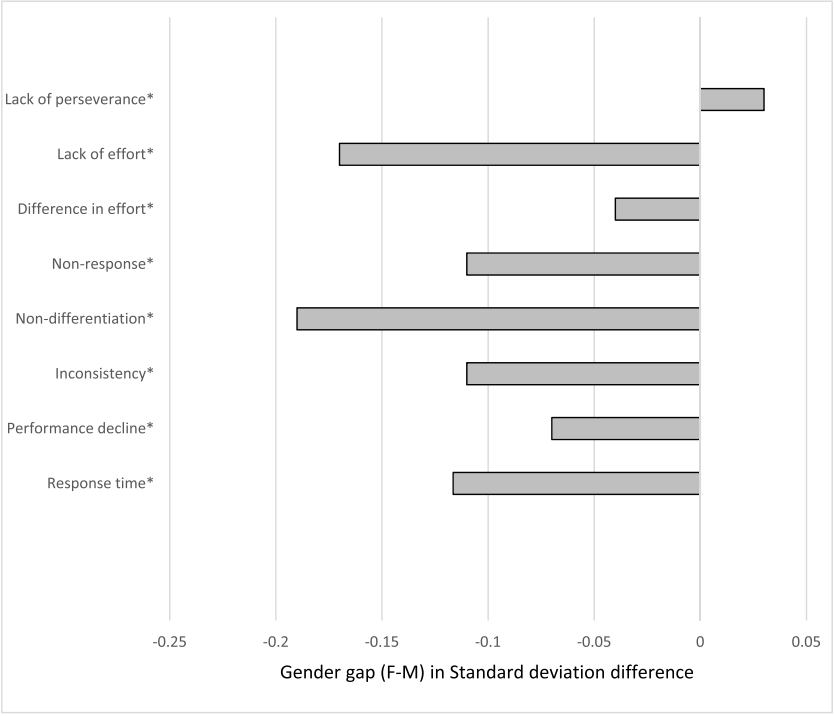


**Fig. 1.** Gender differences in PSEC indicators. Notes: The gender gap refers to the difference between females and males. A negative number implies that females have lower mean values than males. A positive value implies that females have higher mean values than males. An asterisk (*) next to an indicator name indicates that the estimated d value is statistically significant at least at the 5% level. Data for all countries in the sample can be consulted in Table H2 in the Online Supplementary Annex. Source: PISA 2012data.

undifferentiated responses ($d = -0.19$). Among test based behavioural indicators, females displayed, on average, less steep performance decline than males ($d = -0.07$) and were less likely to skip rapidly or guess an answer in the computer-based assessment ($d = -0.12$).

Gender differences varied significantly across countries and indicators. Variability across countries in gender gaps was particularly pronounced for the self-reported lack of perseverance indicator and less pronounced for the performance decline indicator, although a number of outlier countries could be identified. These results suggest that on all behavioural dimensions males appear to have lower PSEC outcomes than females. However, when examining self-reports, males report similar levels of PSEC as females, except for effort expended on the PISA test, indicating an awareness of lower behavioural involvement in the survey. The gender gap in this indicator is similar in size to the behavioural indicators.

Fig. 2 indicates that socio-economically disadvantaged students see themselves as less perseverant than socio-economically advantaged students. On average across countries in our sample the difference between the two groups in the self-reported lack of perseverance indicator corresponds to a medium size gap ($d = -.28$). However, both groups report having expended a similar amount of effort on the PISA test ($d = 0.01$; $p \geq 0.05$). On average, socio-economic differences in behavioural indicators of PSEC exist but are smaller than gaps in self-reported lack of perseverance ($d = -0.12$ for self-reported difference in effort between PISA and a graded PISA test; $d = -0.019$ for item non-response; $d = -0.07$ for non-differentiation; $d = -0.11$ for inconsistency; $d = -0.09$ for performance decline and $d = -0.012$ for response time effort).

Country rankings differ greatly depending on which indicator is considered.

Fig. 3 indicates that there are no differences between students with and those without an immigrant background in self-reported lack of perseverance ($d = -0.03$) and the questionnaire-based behavioural indicator of non-differentiation ($d = 0.01$). However, immigrant students have higher mean values on behavioural indicators based on questionnaire items ($d = 0.18$ for non-response and $d = 0.17$ for inconsistency). The difference between immigrant and non-immigrant students in performance decline is small ($d = 0.06$) and highest for the response time effort indicator ($d = 0.27$).

In many OECD countries students without an immigrant background indicate that they are less perseverant than immigrant students, although on behavioural indicators students with an immigrant background have lower PSEC values than students without such background, a possible reflection of their lower language abilities.

Fig. 3 shows that gaps varied significantly across countries and indicators.

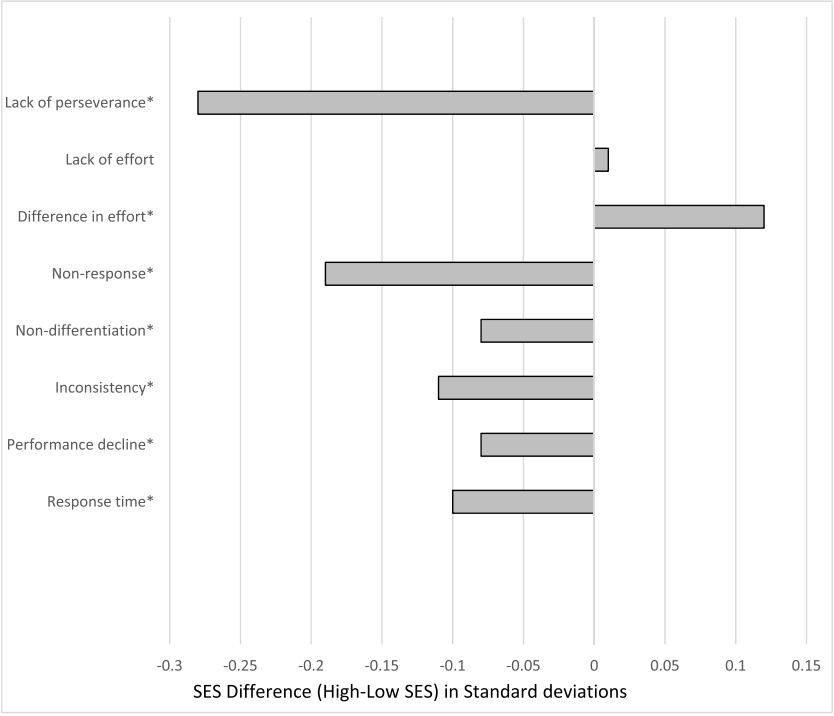


**Fig. 2.** Socio-economic differences in PSEC indicators. Notes: The socio-economic (SES) gap refers to the difference between socio-economically advantaged and socio-economically disadvantaged students. PISA contains an aggregate SES measure based on students' reports on their parents' educational attainment, occupational status and availability of economic and cultural resources in their home. Advantaged students are students in the top 25% of the country specific distribution of SES. Disadvantages students are students in the bottom 25% of the country specific distribution of SES. A negative number implies that advantaged have lower mean values than disadvantaged students. A positive value implies that advantaged have higher mean values than disadvantaged students. An asterisk (*) next to an indicator name indicates that the estimated d value is statistically significant at least at the 5% level. Data for all countries in the sample can be consulted in Table I2 in the Online Supplementary Annex. Source: PISA 2012data.

### *3.3. Relationship with contemporaneous academic achievement*

In Table 3 we report average findings across countries in our sample on the association between each achievement domain in PISA (reading, mathematics, science and problem solving) and each indicators of PSEC. Table 3 suggests that all indicators of PSEC are negatively associated with contemporaneous achievement in reading, mathematics, science and domain general problem solving. Introducing background characteristics generally reduces the association between PSEC indicators and contemporaneous achievement but the reduction is small. On average, across countries in our sample, a change in one SD in the self-reported lack of perseverance is associated with a change of 14% of a *SD* in reading, 17% *SD* in math, 16% *SD* in science and 14% *SD* in problem solving performance. The questionnaire-based behavioural indicator of item non-response is the PSEC measure that is most strongly associated with contemporaneous academic achievement: a change of one SD in non-response is associated with a change of 37% of a *SD* in reading, 33% *SD* in math, 34% *SD* in science and 34% *SD* in problem solving performance. The association between the other questionnaire-based PSEC measures of non-differentiation and inconsistency is smaller: i.e. a change in one *SD* in these measures is associated with a change of between 8% of a *SD* and 19% of a *SD* depending on the achievement domain considered.

Results in Model C of Table 3 suggests that estimated associations between behavioural indicators and achievement are robust to the introduction of self-reported measures (i.e. standardised regression coefficients remain similar in size and statistically significant).

In Fig. 4 we illustrate the average strength of the association between PSEC indicators and mathematics achievement accounting for control variables (as in model B above). Tables for the full sample of countries and the associations of PSEC indicators with all other PISA domains (mathematics, reading, science and problem solving) are available in Tables L1-L4 in the Supplementary Online Annex. Results suggest that associations vary considerably across countries but also for the same country across different indicators. Moreover, relative country rankings depend, to an extent, on the specific indicator used and that the same country can be considered to be above, in line, or below the average depending on the indicator analysed.

### *3.4. Predictive validity of PSEC indicators of outcomes in young adulthood*

Table 4 presents estimates on the relationship between PSEC indicators and upper secondary and tertiary attainment, net of

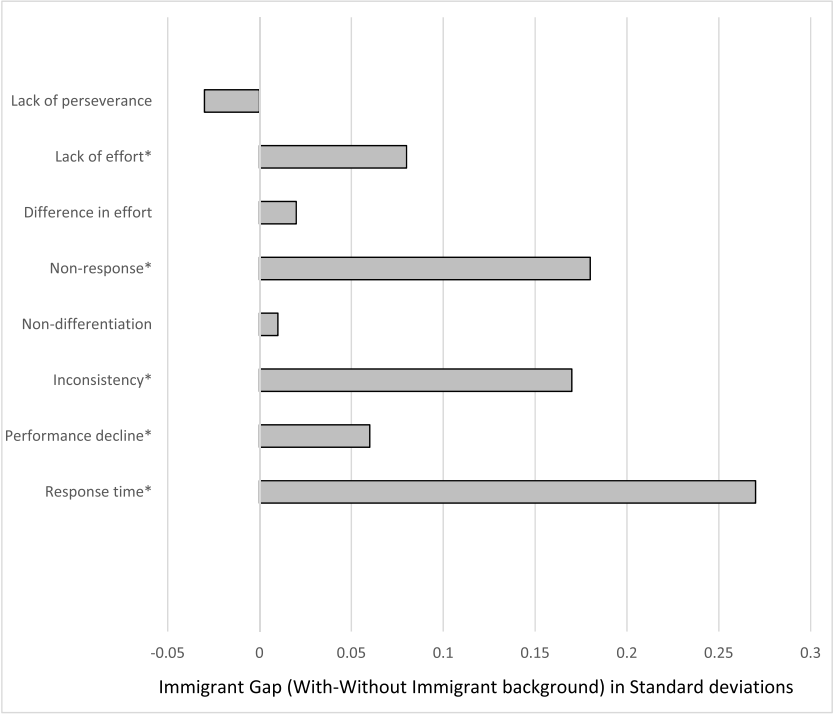


**Fig. 3.** Differences by immigrant background in PSEC indicators. Notes: The immigrant gap refers to the difference between students with and students without an immigrant background. A negative number implies that immigrant students have lower mean values than students without an immigrant background. A positive value implies that immigrant students have higher mean values than students without an immigrant background. An asterisk (*) next to an indicator name indicates that the estimated d value is statistically significant at least at the 5% level. Data for all countries in the sample can be consulted in Table I2 in the Online Supplementary Annex. Source: PISA 2012data.

controls. Coefficients associated with the PSEC indices measure the change in the probability of completing upper-secondary education and university by age 25 as a function of a change in one standard deviation in PSEC indicators.[7] The results indicate that performance decline is associated with university completion in three out of four countries: in Australia and Switzerland, students with similar characteristics and performance who had a one SD greater decline in performance in the PISA test in 2000 had a 7 percentage point lower probability of completing university by age 25 and in Canada they had a 13 percentage points lower probability. By contrast, a difference of one SD in the self-reported lack of perseverance index was associated with a change of 5 percentage points in the probability of completing university by age 25 in Denmark, and 4 percentage points in Switzerland. Among questionnaire-based behavioural indicators, the non-differentiation indicator is not associated with university completion in any of the countries analysed. Inconsistency is associated with university completion in Canada and Switzerland, while non-response is associated with university completion in Denmark and Switzreland and with upper secondary completion in Canada. Indicators of PSEC are less strongly associated with upper secondary school completion than university completion: performance decline is the only indicator that is associated with the probability that individuals will complete upper secondary school in at least two countries (Denmark and Australia).

Table 5 presents results on the relationship between behavioural PSEC indicators and later life outcomes controlling for self-reported PSEC indicators. The table shows that the association between performance decline and educational outcomes at age 25 is statistically significant, quantitatively moderate and robust to the inclusion of self-reported measures of PSEC. Furthermore, models that include the performance decline measure are the ones with the highest explained variance overall and best model fit. We only present the outcome of regressions accounting for the PISA index of perseverance for Denmark and Switzerland and self-reported effort for Australia, but results were similar when we control for any self-reported measure of PSEC and they are available from authors upon request.

## 4. Discussion

This study provides a comprehensive evaluation of the indicators of PSEC that can be derived by examining students' behaviour when completing the PISA cognitive tests and questionnaires. A first contribution of the study consists in demonstrating that these

[7] The standardization of PSEC indicators is based on the pooled SD of each indicator based on 2012 data.

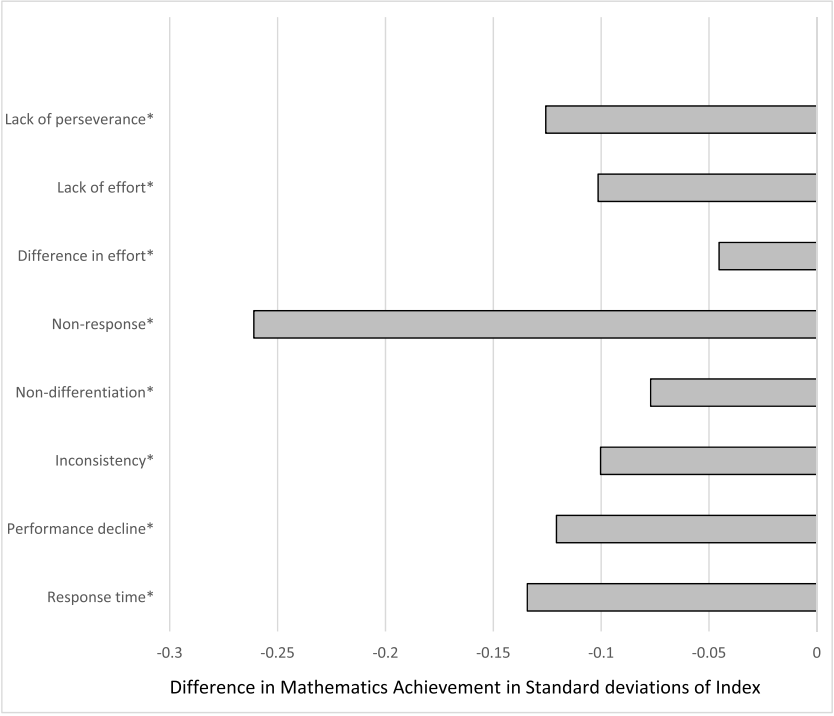


**Fig. 4.** Associations between PSEC indicators and mathematics achievement. Notes: An asterisk (*) next to an indicator name indicates that the estimated coefficient is statistically significant at least at the 5% level. A dark shade indicates that the estimated association for a specific country is statistically significantly above the average for that indicator across the countries with available information ($p \leq 0.05$). Data for all countries in the sample can be consulted in Table K1 in the Online Supplementary Annex and data for other PISA outcomes can be consulted in Tables K2-K4. Source: PISA 2012data.

measures are only limitedly affected by the use of different administration protocols or instruments. Our findings show that country differences in the behavioural measures tend to be relatively stable over successive cycles of the survey.

The results of this study confirm previous findings on the existence of considerable differences in measures of socio-emotional and motivational measures both across and within countries. Behavioural measures indicate that boys have lower levels of PSEC than girls, while boys tend to report that they are more perseverant. Socio-economically disadvantaged students tend to report lower levels of self-reported perseverance, but the gaps between the two groups of students are less marked according to behavioural indicators. These differences across the two sets of measures suggest that self-reports partly reflect subjective judgments that are influenced by gender and socio-economic status.

The behavioural measures of PSEC are also found to be strongly correlated with achievement across all PISA domains, and this correlation is robust to the inclusion of self-reported measures. A novel finding based on longitudinal data from four countries is that some measures of students' behaviour on the test, such as performance decline, are also strongly associated with life outcomes ten years after the test.

Our work suggests that behavioural measures can contribute meaningful additional information to system level monitoring despite the fact that they are constructed as collateral by-products of the testing process. However, these measures should not be considered as replacements for self-report measures. The low correlation between behavioural and self-report measures we document in this study is in line with recent meta-analyses for several constructs (Dang et al., 2020), and suggests that the two sets of measures are not equivalent. The behavioural measures can be conceptualised as measures of "state" because they measure facets of PSEC in the moment and in the specific and highly structured context of test-taking. On the other hand, self-report measures can be interpreted as measures of "trait" PSEC because they ask test participants to reflect on how they usually act and feel across a variety of unstructured, real-life situations (Steinberg and Williams, 2013). If the measurement of traits is favoured by personality theories, such as McClelland's theory (McClelland et al., 1953), other theoretical perspectives such as the social-cognitive perspective (Pintrich et al., 1993) argue for the context-specificity of socio-emotional and motivational constructs, and thus for measuring individuals' states and behaviours in different contexts. Students may be predisposed toward a certain level of persistence (persistence as a trait), yet they may also transiently display different levels of persistence based in different circumstances (persistence as a state). Given the clear conceptual difference between states and traits, self-report and behavioural measures may explain incremental variance above each other, as confirmed in the analyses of predictors of test performance and predictors of education outcomes at age 25 in this paper. On the basis of these findings, we argue that the two sets of measures are complementary and both should thus inform student and country-level monitoring systems.

**Table 4**
The predictive power of PSEC indicators with respect to university and upper secondary school completion.

| Indicator | Outcome | Denmark | | Switzerland | | Australia | | Canada | |
|---|---|---|---|---|---|---|---|---|---|
| | | (N = 1192) | | (N = 1948) | | (N = 3196) | | (N = 5863) | |
| | | b | Adjusted R squared | b | Adjusted R squared | b | Adjusted R squared | b | Adjusted R squared |
| Performance decline | Upper secondary | −0.045* | .0814 | −0.047 | .0611 | −0.0093** | .0496 | −0.027 | 0.039 |
| | University | −0.066 | .148 | −0.074*** | .125 | −0.070* | .205 | −0.132*** | 0.20 |
| Inconsistency | Upper secondary | −0.062 | .0949 | −0.027 | .054 | −0.015 | .051 | −0.009 | 0.036 |
| | University | −0.012 | .139 | −0.031* | .111 | 0.011 | .187 | −0.025* | 0.18 |
| Non-response | Upper secondary | −0.053 | .0811 | 0.023 | .0499 | −0.028 | .0535 | −0.036* | 0.042 |
| | University | −0.050* | .143 | −0.072*** | .115 | −0.014 | .187 | −0.034 | 0.18 |
| Non-differentiation | Upper secondary | 0.0049 | .0729 | 0.0052 | .0487 | −0.012 | .05 | 0.003 | 0.035 |
| | University | 0.015 | .139 | −0.022 | .107 | −0.0015 | .187 | 0.014 | 0.18 |
| Self-reported lack of perseverance | Upper secondary | −0.017 | .0754 | 0.015 | .0512 | | | | |
| | University | −0.049* | .149 | 0.037* | .116 | | | | |
| Self-reported of lack of self-efficacy | Upper secondary | −0.023 | .077 | 0.017 | .0507 | | | | |
| | University | −0.050 | .149 | 0.017 | .108 | | | | |
| Self-reported of lack of instrumental motivation | Upper secondary | 0.0058 | .0731 | 0.026 | .0546 | | | | |
| | University | −0.0040 | .139 | 0.033* | .112 | | | | |
| Self-reported of lack of control expectations | Upper secondary | −0.0052 | .0731 | 0.024 | .0539 | | | | |
| | University | −0.044 | .147 | 0.044** | .117 | | | | |
| Self-reported lack of effort in text | Upper secondary | | | | | −0.016 | .0534 | | |
| | University | | | | | 0.0078 | .187 | | |

Notes: Denmark: PISA 2000-PIAAC 2012 link. Switzerland: PISA 2000-TREE 2010 follow-up. Australia: PISA 2003- LSAY 2013. Canada: PISA 2000-YITS 2010.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

**Table 5**
The incremental predictive power of behavioural PSEC indicators.

| Indicator | Outcome | Denmark | | Switzerland | | Australia | |
|---|---|---|---|---|---|---|---|
| | | (N = 1192) | | (N = 1948) | | (N = 3196) | |
| | | B | Adjusted R squared | B | Adjusted R squared | b | Adjusted R squared |
| Performance decline | Upper secondary | −0.043* | .0823 | −0.046 | .0627 | −0.0074 | .0545 |
| | University | −0.059 | .156 | −0.071*** | .133 | −0.072* | .206 |
| Inconsistency | Upper secondary | −0.061 | .096 | −0.025 | .0553 | −0.014 | .0563 |
| | University | −0.0088 | .149 | −0.025 | .118 | 0.011 | .187 |
| Non-response | Upper secondary | −0.050 | .0821 | 0.027 | .0527 | −0.026 | .0587 |
| | University | −0.042 | .151 | −0.063** | .122 | −0.015 | .187 |
| Non-differentiation | Upper secondary | 0.0047 | .0747 | 0.0096 | .051 | −0.012 | .056 |
| | University | 0.014 | .149 | −0.012 | .116 | −0.0013 | .187 |

Notes: Denmark: PISA 2000-PIAAC 2012 link. Switzerland: PISA 2000-TREE 2010 follow-up. Australia: PISA 2003- LSAY 2013. Canada: PISA 2000-YITS 2010.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Another possible explanation for the low correlation between the measures reviewed in this study is measurement error. There is evidence that behavioural measures have relatively large trial-by-trial variations even in experiment studies (Dang et al., 2020; Rouder and Haaf, 2019). One limitation of this study is that the cross-sectional nature of the data only allows investigating the stability of the measures at the country-level, while a more comprehensive analysis of reliability would require taking repeated observations for the same students.

A further limitation of this study is that the evidence argument to support the use of the behavioural measures of PSEC for comparing countries is fully based on their correlation with other measures of PSEC (self-report measures) and with measures of educational outcomes (performance in the PISA test and completion of higher education). Given the relatively weak theoretical foundations of these measures, it would be important to develop other types of validity evidence. Further validation could be based on

small-scale, multi-modal studies where the evidence on engagement from the log data is triangulated with other evidence based on direct observation of students' facial or body expressions during the test, self-reports from concurrent and retrospective interviews, or with evidence from eye-tracking or electroencephalogram (EEG) devices (Apicella et al., 2022).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssresearch.2023.102874.

## References


Akyol, S.P., Krishna, K., Wang, J., 2018. Taking PISA Seriously: How Accurate Are Low Stakes Exams? NBER Working Paper No. 24930. NBER, Cambridge.

Allom, V., Panetta, G., Mullan, B., Hagger, M.S., 2016. Self-report and behavioural 4 approaches to the measurement of self-control: are we assessing the same 5 construct? Pers. Indiv. Differ. 90, 137–142. https://doi: 6 10.1016/j.paid.2015.10.051.

Almlund, M., Duckworth, A.L., Heckman, J., Kautz, T., 2011. Personality psychology and economics. In: Hanushek, E.A., Machin, S., Woessmann, L. (Eds.), Handbook of the Economics of Education, vol. 4, pp. 1–181. https://doi.org/10.1016/B978-0-444-53444-6.00001-8.

Apicella, A., Arpaia, P., Isgrò, F., Mastrati, G., Moccaldi, N., 2022. A survey on EEG-based solutions for emotion recognition with a low number of channels. In: IEEE Access, pp. 117411–117428. https://doi.org/10.1109/ACCESS.2022.3219844, 10.

Balart, P., Oosterveen, M., 2019. Females show more sustained performance during test-taking than males. Nat. Commun. 10 (1) https://doi.org/10.1038/s41467-019-11691-y.

Barge, S., Gehlbach, H., 2012. Using the theory of satisficing to evaluate the quality of survey data. Res. High. Educ. 53 (2), 182–200. https://doi.org/10.1007/s11162-011-9251-2.

Barry, C.L., Horst, S.J., Finney, S.J., Brown, A.R., Kopp, J.P., 2010. Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. Int. J. Test. 10 (4), 342–363. https://doi.org/10.1080/15305058.2010.508569.

Borghans, L., Schils, T., 2013. The Leaning Tower of Pisa: Decomposing Achievement Test Scores into Cognitive and Noncognitive Components. Unpublished manuscript. Draft version: July 22, 2013.

Borgonovi, F., 2021. Is the literacy achievement of teenage boys poorer than that of teenage girls, or do estimates of gender gaps depend on the test? A comparison of PISA and PIAAC. J. Educ. Psychol. https://doi.org/10.1037/edu0000659.

Borgonovi, F., Biecek, P., 2016. An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. Learn. Indiv Differ 49, 128–137.

Brunello, G., Crema, A., Rocco, L., 2018. Testing at Length if it Is Cognitive or Non-cognitive. Institute of Labor Economics (IZA), N. 11603.Bonn. IZA Discussion Papers.

Cohn, A., Maréchal, M.A., Tannenbaum, D., Zünd, C.L., 2019. Civic honesty around the globe. Science 365 (6448), 70–73. https://doi.org/10.1126/science.aau8712.

Cole, J.S., Bergin, D.A., Whittaker, T.A., 2008. Predicting student achievement for low stakes tests with effort and task value. Contemp. Educ. Psychol. 33 (4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002.

Credé, M., Tynan, M.C., Harms, P.D., 2017. Much ado about grit: a meta-analytic synthesis of the grit literature. J. Pers. Soc. Psychol. 113 (3), 492–511. https://doi.org/10.1037/pspp0000102.

Cyders, M.A., Coskunpinar, A., 2011. Measurement of constructs using self-report and behavioral lab tasks: is there overlap in nomothetic span and construct representation for impulsivity? Clin. Psychol. Rev. 31 (6), 965–982.

Danner, D., Lechner, C.M., Spengler, M., 2021. Do we need socio-emotional skills? Front. Psychol. 6 (12), 723470 https://doi.org/10.3389/fpsyg.2021.723470.

Dang, J., King, K.M., Inzlicht, M., 2020. Why are self-report and behavioral measures weakly correlated?Trends in Cognitive Sciences 24 (4), 267–269. https://doi.org/10.1016/j.tics.2020.01.007.

Debeer, D., Janssen, R., 2013. Modeling item-position effects within an IRT framework. J. Educ. Meas. 50 (2), 164–185. https://doi.org/10.1111/jedm.12009.

Duckworth, A., et al., 2007. Grit: perseverance and passion for long-term goals. J. Pers. Soc. Psychol. 92 (6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087.

Duckworth, A.L., Seligman, M.E.P., 2005. Self-discipline outdoes IQ in predicting academic performance of adolescents. Psychol. Sci. 16 (12), 939–944. https://doi.org/10.1111/j.1467-9280.2005.01641.x.

Duckworth, A.L., Kern, M.L., 2011. A meta-analysis of the convergent validity of self-control measures. J. Res. Pers. 45, 259–268.

Duckworth, A.L., Yeager, D.S., 2015. Measurement matters: assessing personal qualities other than cognitive ability for educational purposes. Educ. Res. 44 (4), 237–251. https://doi.org/10.3102/0013189X15584327.

Duckworth, A., Quinn, P., 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. J. Educ. Psychol. 104 (2), 439–451. https://doi.org/10.1037/a0026280.

Fahle, E.M., Lee, M.G., Loeb, S., 2019. A Middle School Drop: Consistent Gender Differences in Students' Self-Efficacy. Policy Analysis for California Education, PACE. Working Paper.

Falk, A., Hermle, J., 2018. Relationship of gender differences in preferences to economic development and gender equality. Science 362 (6412). https://doi.org/10.1126/science.aas9899.

Fiske, D.W., 1971. Measuring the Concepts of Personality. Aldline Publishing Co, Chicago, Illinois.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., Klieme, E., 2014. The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. J. Educ. Psychol. 106 (3), 608–626.

Gneezy, U., List, J.A., Livingston, J.A., Qin, X., Sadoff, S., Xu, Y., 2019. Measuring success in education: the role of effort on the test itself. Am. Econ. Rev. 1 (3), 291–308.

Greiff, S., Borgonovi, F., 2022. Teaching of 21st century skills needs to be informed by psychological research. Nature Reviews Psychology 1, 314–315. https://doi.org/10.1038/s44159-022-00064-w.

Gutman, L.M., Schoon, I., 2016. A synthesis of causal evidence linking non-cognitive skills to later outcomes for children and adolescents. In: Khine, M.S., Areepattamannil, S. (Eds.), Non-cognitive Skills and Factors in Educational Attainment, pp. 171–198. https://doi.org/10.1007/978-94-6300-591-3_9.

Heckman, J.J., Kautz, T., 2012. Hard evidence on soft skills. Lab. Econ. 19 (4), 451–464. https://doi.org/10.1016/j.labeco.2012.05.014.

Heckman, J.J., Pinto, R., Savelyev, P., 2013. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. Am. Econ. Rev. 103 (6), 2052–2086. https://doi.org/10.1257/aer.103.6.2052.

Heckman, J.J., Stixrud, J., Urzua, S., 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. J. Labor Econ. 24 (3), 411–482. https://doi.org/10.1086/504455.

Hitt, C., 2015. *Just Filling in the Bubbles: Using Careless Answer Patterns on Surveys as a Proxy Measure of Noncognitive Skills* EDRE Working Paper 2015-06. http://www.uaedreform.org/downloads/2015/07/edre-working-paper-2015-06.pdf.

Hitt, C., Trivitt, J., Cheng, A., 2016. When you say nothing at all: the predictive power of student effort on surveys. Econ. Educ. Rev. 52, 105–119. https://doi.org/10.1016/j.econedurev.2016.02.001.

Kane, M., 2006. Content-related validity evidence in test development. In: Haladyna, T.M., Downing, S.M. (Eds.), Handbook of Test Development. Routledge, New York. https://doi.org/10.4324/9780203874776.
Kane, M., 2013. Validating the interpretations and uses of test scores. J. Educ. Meas. 50 (1), 1–73. https://doi.org/10.1111/jedm.2013.50.issue-1.
Kankaraš, M., 2017. Personality matters: relevance and assessment of personality characteristics. In: OECD Education Working Papers 157. OECD Publishing, Paris. https://doi.org/10.1787/8a294376-en.
Kautz, T., Heckman, J.J., Diris, R., Borghans, L., 2014. Fostering and measuring skills: improving cognitive and non-cognitive skills to promote lifetime success. In: OECD Education Working Papers 110. OECD Publishing, Paris.
Knowles, E.S., 1988. Item context effects on personality scales: measuring changes the measure. J. Pers. Soc. Psychol. 55 (2), 312–320. https://doi.org/10.1037/0022-3514.55.2.312.
Kroehne, U., Goldhammer, F., 2018. How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. Behaviormetrika 45 (2), 527–563. https://doi.org/10.1007/s41237-018-0063-y.
Krosnick, J.A., 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl. Cognit. Psychol. 5 (3), 213–236. https://doi.org/10.1002/acp.2350050305.
Kuhfeld, M., Soland, J., 2020. Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. Journal of Research on Educational Effectiveness 13 (1), 147–175. https://doi.org/10.1080/19345747.2019.1636437.
Kyllonen, P., Kell, H., 2018. Ability tests measure personality, personality tests measure ability: disentangling construct and method in evaluating the relationship between personality and ability. J. Intell. 6 (3), 32. https://doi.org/10.3390/jintelligence6030032.
Local Burden of Disease Educational Attainment Collaborators, 2020. Mapping disparities in education across low- and middle-income countries. Nature 577, 235–238. https://doi.org/10.1038/s41586-019-1872-1.
McClelland, D.C., Atkinson, J., Clark, R., Lowell, E., 1953. The Achievement Motive. Appleton-Century-Crofts, New York.
Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R., et al., 2001. Psychological testing and psychological assessment: a review of evidence and issues. Amercian Psychologist 56, 128–165.
Mood, C., 2010. Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur. Socio Rev. 26 (1), 67–82. https://doi.org/10.1093/esr/jcp006.
Organisation for Economic Co-operation and Development OECD, 2009. PISA Data Analysis Manual: SAS, second ed. OECD Publishing, Paris.
OECD, 2014. PISA 2012 Technical Report. OECD Publishing, Paris.
OECD, 2021. Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills. OECD Publishing, Paris. https://doi.org/10.1787/92a11084-en.
OECD, 2022. Education at a Glance. OECD Publishing, Paris.
Ones, D.S., Viswesvaran, C., Reiss, A.D., 1996. Role of social desirability in personality testing for personnel selection: the red herring. J. Appl. Psychol. 81 (6), 660–679.
Pintrich, P.R., Marx, R.W., Boyle, R.A., 1993. Beyond cold conceptual change: the role of motivational beliefs and classroom contextual factors in the process of conceptual change. Rev. Educ. Res. 63, 167–199. https://doi.org/10.3102/00346543063002167.
Poropat, A., 2009. A meta-analysis of the five-factor model of personality and academic performance. Psychol. Bull. 135 (2), 322–338. https://doi.org/10.1037/a0014996.
Revelle, W., 2007. Experimental approaches to the study of personality. In: Robins, R.W., Fraley, R.C., Krueger, R.F. (Eds.), Handbook of Research Methods in Personality Psychology. The Guilford Press, pp. 37–61.
Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., Goldberg, L.R., 2007. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspect. Psychol. Sci. 2 (4), 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x.
Rosander, P., Bäckström, M., 2014. Personality traits measured at baseline can predict academic performance un upper secondary school three years later. Personality and Social Psychology 55, 611–618. https://doi.org/10.1111/sjop.12165.
Rouder, J.N., Haaf, J.M., 2019. A psychometrics of individual differences in experimental tasks. Psychological Bulletin Review 26, 452–467. https://doi.org/10.3758/s13423-018-1558-y.
Salganik, M., 2019. Bit by Bit: Social Research in the Digital Age. Princeton University Press, Princeton.
Sharma, L., Markon, K.E., Clark, L.A., 2014. Toward a theory of distinct types of "impulsive" behaviors: a meta-analysis of self-report and behavioral measures. Psychol. Bull. 140 (2), 374–408. https://doi.org/10.1037/a0034418.
Soland, J., 2018. The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. Appl. Meas. Educ. 31 (4), 312–323.
Soland, J., Zamarro, G., Cheng, A., Hitt, C., 2019. Identifying naturally occurring direct assessments of social-emotional competencies: the promise and limitations of survey and assessment disengagement metadata. Educ. Res. 48 (7), 466–478.
Soland, J., 2019. Can item response times provide insight into students' motivation and self-efficacy in math? An initial application of test metadata to understand students' social–emotional needs. Educ. Meas. 38 (3), 86–96.
Soland, J., Kuhfeld, M., 2019. Do students rapidly guess repeatedly over time? A longitudinal analysis of student test disengagement, background, and attitudes. Educ. Assess. 24 (4), 327–342.
Soto, C.J., 2019. How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project. Psychol. Sci. 30 (5), 711–727.
Soto, C.J., 2020. Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. Soc. Psychol. Personal. Sci. https://doi.org/10.1177/1948550619900572.
Steinberg, M.L., Williams, J.M., 2013. State, but not trait, measures of persistence are related to negative affect. J. Stud. Alcohol Drugs 74 (4), 584–588. https://doi.org/10.15288/jsad.2013.74.584.
Vannette, D.L., Krosnick, J.A., 2014. Answering questions: a comparison of survey satisficing and mindlessness. In: Ie, A., Ngnoumen, C.T., Langer, E.J. (Eds.), The Wiley Blackwell Handbook of Mindfulness. John Wiley & Sons, pp. 312–327. https://doi.org/10.1002/9781118294895.ch17.
Wilmot, M.P., Ones, D.S., 2019. A century of research on conscientiousness at work. Proc. Natl. Acad. Sci. USA 116 (46), 23004–23010.
Wise, S., 2017. Rapid-guessing behavior: its identification, interpretation, and implications. Educ. Meas. 36 (4), 52–61. https://doi.org/10.1111/emip.12165.
Wise, S.L., DeMars, C.E., 2006. An application of item response time: the effort-moderated IRT model. J. Educ. Meas. 43 (1), 19–38.
Wise, S.L., DeMars, C.E., 2010. Examinee noneffort and the validity of program assessment results. Educ. Assess. 15 (1), 27–41. https://doi.org/10.1080/10627191003673216.
Wise, S.L., Kingsbury, G., 2002. Performance decline as an indicator of generalized test-taking disengagement. Appl. Meas. Educ. https://doi.org/10.1080/08957347.2022.2155651.
Wise, S.L., Kong, X., 2005. Response time effort: a new measure of examinee motivation in computer-based tests. Appl. Meas. Educ. 18, 163–183. https://doi.org/10.1207/s15324818ame1802_2.
Wise, S.L., Ma, L., 2012. Setting response time thresholds for a CAT item pool: the normative threshold method. In: Annual Meeting of the National Council on Measurement in Education. Vancouver, Canada. https://nwea.org/content/uploads/2012/04/Setting-Response-Time-Thresholds-for-a-CAT-Item-Pool.pdf.
Zamarro, G., Cheng, A., Shakeel, M.D., Hitt, C., 2018. Comparing and validating measures of non-cognitive traits: performance task measures and self-reports from a nationally representative internet panel. Journal of Behavioral and Experimental Economics 72, 51–60. https://doi.org/10.1016/j.socec.2017.11.005.
Zamarro, G., Hitt, C., Mendez, I., 2019. When students don't care: reexamining international differences in achievement and student effort. J. Hum. Cap. 13 (4), 519–552. https://doi.org/10.1086/705799.