Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

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

Measuring non-cognitive skills exploiting log-files on online behaviour

Yi Zhang [a,b,*], Jia He [b,c]

[a] China Center for Human Capital and Labor Market Research, Central University of Finance and Economics, 39 South College Road, Haidian District, Beijing, P.R. China, 100081
[b] Tilburg University, the Netherlands
[c] German Institute for International Educational Research, Germany

ARTICLE INFO

JEL Codes:
D91 I21 J24 N30

Keywords:
Non-cognitive skill
Self-reported measure
Behavioural measure
Log-file
Comparability
Predictive performance

ABSTRACT

Non-cognitive skills are important components of individuals' human capital and good predictors of educational and labour market outcomes. Conventional self-reported measures of non-cognitive skills suffer from measurement errors stemming from self-presentation and reference group effects, which can produce paradoxical results in cross-country comparisons. We propose a novel source of measures derived from computer-generated log files on the behaviour of individuals taking an online test or respondents taking an online survey. We analyse measures of two desirable non-cognitive skills, perseverance and deep learning, constructed with log-file data from two large-scale educational surveys. Compared with the self-reported measures, our log-based behavioural measures have higher cross-country comparability, as they predict the performance of tests consistently at both individual and country levels. They also show high predictive validity in schooling and labour market outcomes, offering promise for a wide range of applications. We discuss the methodological implications of log-based behavioural measures and encourage researchers to apply them in combination with conventional self-reported measures.

1. Introduction

Non-cognitive skills, such as social skills, perseverance, and deep learning, are important components of individuals' human capital. They appear to be good predictors of schooling and career success (see, e.g., Weiss, 1988; Heckman et al., 2006; Duckworth et al., 2007). Moreover, their malleability in early life opens the door for interventions through education policy (García, 2016; West et al., 2016).

On the other hand, measuring non-cognitive skills is challenging. Conventional measures using Likert scale self-reports are criticised for the lack of measurement comparability, in both the economics and the psychology literature.[1] Bond and Lang (2019) point out that measures reported in ordered intervals (e.g., happiness) are only comparable across groups under a rather strong assumption that all respondents share a common reporting scale. Similarly, Van de Gaer et al. (2012) show that self-reported Likert-scale measures are not necessarily comparable in cross-cultural analysis due to heterogeneous reporting behaviour. A classic example comes from the motivation-achievement paradox. With data from the Programme for International Student Assessment (PISA), students' self-reported learning motivation is often found to be positively related to academic achievement within each participating country. However, when scores are aggregated at the country level and the correlation is computed between countries' average levels of motivation and achievement, a negative correlation is found. For example, East Asian countries, such as China, Korea, and Japan typically show *high* scores on achievement in PISA studies, but tend to have *low* scores on learning motivation. Such a paradox is partially attributable to the reference group effect, implying that respondents use different implicit standards (influenced by their immediate social context) in their self-evaluations, or due to their different styles of presenting themselves (e.g., response amplification through the tendency to endorse the end points of a scale, or response moderation through the endorsement of the midpoint of a scale).

Various strategies have been proposed to alleviate measurement incomparability of self-reported Likert-scale measures. Correction procedures such as anchoring vignettes and alternative item formats such as forced-choice responses are employed to enhance their comparability (e.g., Kapteyn et al., 2007; Voňková & Hullegie, 2011; Kyllonen &

* Corresponding author.
E-mail address: y.zhang_3@hotmail.com (Y. Zhang).

[1] A related literature points out that self-reported measures of noncognitive skills can be sensitive to survey administration conditions. See, for example, Chen et al. (2020) for a nice discussion.

https://doi.org/10.1016/j.socec.2023.101992
Received 6 March 2022; Received in revised form 14 January 2023; Accepted 9 February 2023
Available online 12 February 2023

Bertling, 2014; Leising et al., 2015; Robert et al., 2015). Assessment by a third party (e.g. peers, guardians, teachers) can be used to mitigate the effect of self-presentation styles (e.g., Konstabel et al., 2006, Feng et al., 2022). But correction procedures are usually unavailable for measuring non-cognitive skills, while third-party assessments are limited to relatively small-scale studies.

Another strand of the literature explores alternative measures of non-cognitive skills, such as behavioural measures. Heckman and Rubinstein (2001) use a behavioural indicator (receiving the "General Educational Development" testing program), as a proxy for (low) non-cognitive skills. Lindqvist and Vestman (2011) utilise administrative records of suitability assessment for military service as a measure of non-cognitive skills. Hitt et al. (2016) and Zamarro et al. (2018) use survey-effort measures (e.g. item nonresponse rates, careless answering) to proxy non-cognitive skills related to conscientiousness and neuroticism. Other measures include behavioural observation and coding of survey respondents by interviewers or experimenters (e.g., Renninger & Bachrach, 2015) and measuring non-cognitive skills on the basis of task performance (e.g., Reynolds et al., 2006). Though these behavioural measures can be context-specific and more sensitive to incentives and situational factors (Lundberg, 2015), yet they have clear advantages over self-reports as being more objective, and therefore are less plagued by incomparable reporting styles.

We add to the literature by proposing a new source of behavioural measures: we propose to use computer-generated logs to construct behavioural indicators for non-cognitive skills. In computer-based assessments (such as the well-known Programme for International Student Assessment, i.e., PISA test), log files record respondents' sequences of actions like keystrokes and mouse clicks etc., from which we extract behavioural measures to quantify certain non-cognitive skills. The objective and unobtrusive nature of such measures makes them immune to respondents' self-presentation styles or to reference group effects. They thus hold promise in validating self-report data and predicting achievement in a cross-group (esp. cross-cultural) context. As log-files are increasingly available in large-scale international online tests or surveys, compared to the existing behavioural measures based on lab tasks, log-based behavioural measures can reach a larger population and potentially be applied to more dimensions of non-cognitive skills.

We analyse two examples of non-cognitive skills that are considered important to education policy: perseverance and deep learning (García, 2016). We use the log files from the Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC) to construct our behavioural indicators for the two skills, respectively. We show that log-based behavioural measures have higher cross-country comparability than self-assessments, as they predict the performance of tests consistently at individual and country levels, while conventional self-reported measures do not. We further show that log-based behavioural measures have high predictive validity in forecasting individuals' schooling and labour market outcomes, indicating that the log-based behavioural measure holds promise for a wide range of applications. Our main results are robust to an alternative data-driven interpretation of behavioural measures and insensitive to alternative model specifications. We also discuss the methodological implications of log file-based behavioural measures, and encourage researchers to apply them in combination with conventional self-reported measures.

2. Two examples: perseverance and deep learning

2.1. *Example 1: perseverance in PISA*

2.1.1. *Data*

2.1.1.1. Data source. We use the log data and background questionnaire of PISA in 2012. The PISA test has a computer-based assessment targeting 15-year-old students in 42 countries. It includes a cognitive test on math, digital reading, and problem solving, and a background questionnaire on various attitudes and behaviours related to learning. Data on the background questionnaire, cognitive tests, and log files of sampled cognitive items are published for public research use on the OECD website (OECD, 2013a, 2013b). The target construct is perseverance in both the questionnaire and from computer generated log-files in the cognitive assessment.

2.1.1.2. Sample restriction. To compare with our log-based behavioural measure, we need a valid and comparable self-reported measure of perseverance from Likert-scale items of perseverance. To ensure that this self-reported measure captures the same construct across countries, we perform a multi-group confirmatory factor analysis and country-wise internal consistency checks. We find that data on perseverance show a lack of construct equivalence in the United Arab Emirates, Brazil, Bulgaria, Columbia, Hungary, Malaysia, Montenegro, Slovenia and Serbia. We therefore drop observations from these 9 countries, retaining a sample of 33 countries and 14,888 observations. Online Appendix A provides details on the multi-group confirmatory factor analysis and the sample restriction.

2.1.2. *Measures*

2.1.2.1. Log file-based behavioural measure of perseverance. Perseverance is defined as "one's tendency to persist and endure in the face of adversity" (Eisenberger 1992; Markman et al. 2005). Studies on lab task-based behavioural measures typically use "time spent in a difficult task" or "the number of attempted (impossible) trials" to measure perseverance (Määttänen et al. 2021). Following this line of thought, we look for acts recorded in log-files that could reflect the number of persistent trials toward goals in spite of frustrations. The number of clicks of the "RESET" button in the "Traffic" unit in the problem-solving cognitive assessment could fit this idea.

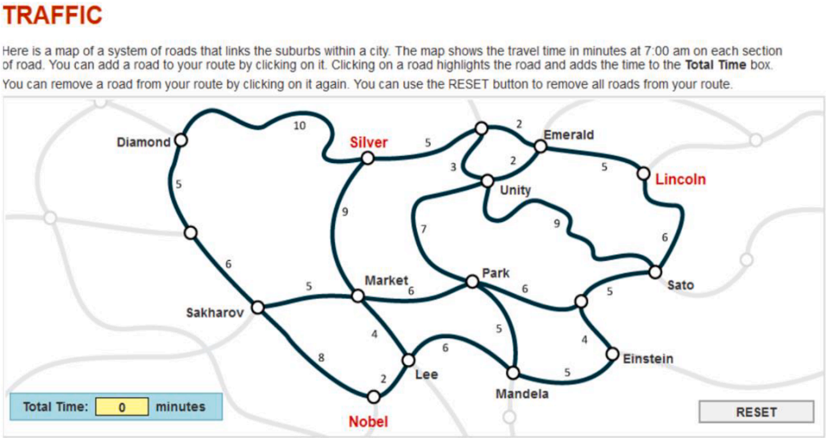
The "Traffic" unit provides a map connecting different areas with the following description:

> "*Here is a map of a system of roads that links the suburbs within a city. The map shows the travel time in minutes at 7:00 am on each section of road. You can add a road to your route by clicking on it. Clicking on a road highlights the road and adds the time to the Total Time box. You can remove a road from your route by clicking on it again. You can use the RESET button to remove all roads from your route.*"

The map can be found in Fig. 1. Respondents are asked to utilise the interactive map to answer three questions: a calculation of the time needed for the shortest route between two specific areas (Question 1), highlighting the shortest route from two areas on the map (Question 2), and selecting the best place for three persons living in different areas to meet, given that no one needs to travel more than 15 min (Question 3).

The answers to these questions are not obvious without trying different routes. By clicking RESET, one keeps finding answers in spite of failed trials (the frustration). The respondent's trials and errors until reaching the answer naturally reflect the extent to which the respondent does not give up and persists in pursuing their goal. Perseverance here is therefore operationalised as the total number of trials and errors by clicking the RESET button – we extract the number of resets for each question and add them up to obtain the behavioural indicator of perseverance for each individual. The average ranges from 1.98 in Croatia to 6.23 in South Korea.[2]

[2] One concern is that too many resets might capture noise rather than perseverance. In Section 3.2, we trim reset clicks greater than the 95th percentile and show that results are robust.



TRAFFIC

Here is a map of a system of roads that links the suburbs within a city. The map shows the travel time in minutes at 7:00 am on each section of road. You can add a road to your route by clicking on it. Clicking on a road highlights the road and adds the time to the **Total Time** box.

You can remove a road from your route by clicking on it again. You can use the RESET button to remove all roads from your route.

Total Time: 0 minutes

Source: http://www.oecd.org/pisa/pisaproducts/pisa2012problemsolvingquestions.htm

Fig. 1. Screenshot of the “Traffic unit” map.

2.1.2.2. Self-reported perseverance. We extract a continuous measure with factor analysis from 4 perseverance-related items.[3] These items are: “When confronted with a problem I give up easily”, “I remain interested in the tasks that I start”, “I continue working on tasks until everything is perfect”, and “When confronted with a problem I do more than what is expected of me.” They all have the same response options ranging from 1 (*very much like me*) to 5 (*not at all like me*). The value of self-reported perseverance ranges from −2.26 to 1.67, with a mean of 0 and a standard deviation of 0.90.[4] A larger value indicates a higher level of perseverance.

Note that here, we interpret “the number of resets” in a top-down manner. That is, we start from the definition of perseverance, and find the log-based behaviour that fits the definition. Then we compare it directly with the self-reported perseverance. In Section 3.1, we explore a data-driven approach for interpretation. We check what “the number of resets” reflects by correlating it with various dimensions of self-reported traits. Picking the most relevant dimensions, we then use principal component analysis (PCA) to construct a one-dimensional self-report for comparison.

2.1.2.3. Outcome variable. **Traffic unit performance:** The total grades for three questions in the “Traffic unit”. The value ranges from 0 to 3, one point for each correct answer, with a mean score of 2.38 and a standard deviation of 0.83. We use our measures of perseverance to predict this Traffic unit score. We expect perseverance, as a desirable non-cognitive skill for the learning process, to be positively correlated with this cognitive test performance.

2.1.2.4. Control variables. **Maternal education:** a larger total number of resets may reflect a lower innate ability instead of measuring a higher level of perseverance. We use maternal levels of education (“upper secondary-academic”, “upper secondary-vocational”, “lower secondary”, “primary level”, and “not finish primary level”) to proxy the respondent’s innate ability.[5] 55% of the mothers completed the academic upper secondary level of education.

Having a computer at home or not: The total number of resets might capture acquaintance with information and communications technology (ICT) besides perseverance. We therefore control for having a computer at home or not to proxy for the ICT proficiency. 93% of the respondents have computers at home.

We further control for gender (50% of the sample are girls), country fixed effects, and birth year fixed effects in the prediction exercises. Note that the PISA test is for children aged 15, so the respondents in our sample were born in either 1996 or 1997 (92% born in 1996).[6]

2.1.3. Descriptive statistics

As two different measures of perseverance, we expect the behavioural and self-reported measures to show some consistency. Indeed as seen in Table 1, the two measures are positively though only modestly correlated ($\rho_{self,behav}$=0.017) at the individual level.

At the individual level, both measures of perseverance are positively correlated with the cognitive test performance, as expected ($\rho_{self,score}$

Table 1
Correlation matrix of test performance and perseverance.

Variable	Traffic unit Performance	Self-reported perseverance
Traffic unit Performance	–	
Self-reported perseverance	0.0321***	–
Behavioural perseverance measure	0.0732***	0.0173**

Note: * Significant at 10%; ** at 5%; *** at 1%. Sample size 14,826.

[3] We do not include the item “I put off difficult problems” because it is found to be cross-country incomparable in the metric equivalence check. See Online Appendix A.

[4] The standard deviation would be 1 if we extract the factor scores in a single CFA model, but we do it in the metric invariance model of the multigroup CFA, where each country gets it is own mean of 0 and SD of 1, and pooling them together, the SD is slightly deviating from 1.

[5] Note that better measures for innate ability would be standardized test scores of cognitive abilities, e.g., intelligence quotient (IQ). Such information is not available in PISA or PIAAC. We therefore use parental education as a proxy for pre-determined cognitive abilities. In Section 3.2, we further control for “friends’ performance in math” for a robustness check, to at least partly account for one’s innate cognitive ability.

[6] In Section 3.2, we include *wealth, immigration status, household structure, education resources at home*, and *friends’ math performance* as additional controls. The results are insensitive to further controls. We therefore keep a parsimonious specification for the main analysis.

=0.032 for the self-reported measure and $\rho_{\text{behav,score}}$ =0.073 for the behavioural measure).

However, when aggregated to the country level, self-reported perseverance shows a paradoxical negative correlation with the mean performance (Fig. 2(A)). Students from East Asian countries and regions like Japan (JPN) and Chinese Taipei (TAP), generally known as perseverant and modest, tend to self-report a lower perseverance score. Just like the "motivation-achievement" paradox, this "perseverance-achievement" paradox suggests that self-reported measures can be plagued by reference-group effects or group-specific reporting styles, especially in cross-group analyses. In contrast, Fig. 2(B) shows that the behavioural measure of perseverance is less prone to heterogeneous reporting behaviour, and retains the positive correlation in cross-country comparisons.

2.1.4. *Predictive performance on cognitive test scores*

We further compare the predictive performance of both measures on "Traffic unit" test scores controlling for possible confounders such as innate ability, ICT proficiency, and demographic characteristics. A good measure of perseverance is expected to have a positive partial correlation with test scores both at the individual level and at the country level.

Table 2 reports OLS estimates of linear regression models at the individual level. Columns (1) and (2) both show a positive partial correlation between two measures of perseverance and test scores, controlling for individual characteristics. Columns (3) to (5) control for both perseverance measures in the same regression. Adding other controls or not, both measures strongly associate with test scores, indicating the complementarity between two measures. In Column (5) with standardised perseverance measures, one standard deviation increase in self-reported and behavioural perseverance will significantly improve the test score by a similar magnitude.

In short, at the individual level, both self-reported and log file-based behavioural measures of perseverance predict test performance well with some complementarity.

We then aggregate all variables to the country level, i.e., each observation is a country average of the variable. We perform similar regressions as in Table 2 at the country level and report OLS estimates in Table 3. Similar to the "perseverance-achievement" paradox in Fig. 2 (A), in columns (1) and (2), a higher self-reported perseverance is negatively associated with test performance. And this paradoxical negative correlation becomes even larger if we add controls for the behavioural perseverance measure (columns (5) to (7)). In comparison, the unobtrusive objective behavioural measure consistently shows a positive association with the test performance in all specifications, indicating that the log file-based behavioural measure is especially recommended in cross-group comparison settings.

2.2. *Example 2: deep learning in PIAAC*

To show that the predictive advantage of our proposed log-based behavioural measure is not merely by chance, we construct behavioural measures for another desirable non-cognitive skill, namely deep learning, using a different data source and perform similar predictive exercises as in Example 1.

2.2.1. *Data*

2.2.1.1. Data source. We use the data of log files and cognitive test scores from PIAAC 2013. PIAAC has a computer-based assessment in 24 countries targeting working adults aged between 16 and 65 years old. Data from the background questionnaire, cognitive test scores, and log file data of the cognitive assessment in 16 countries are published for research use (OECD, 2017).

2.2.1.2. Sample restriction. Given the booklet design (planned missing) in the cognitive assessment and the availability of specific unit log file information, we restricted our analysis to respondents in 13 countries with at least one valid response in the targeted log files of problem-solving tasks.[7] These countries are Austria, Belgium, Germany, Estonia, Finland, Ireland, Netherlands, Norway, Poland, Slovak Republic, UK, and USA. The resulting total sample size is 20,167 observations.

2.2.2. *Measures*

2.2.2.1. Two log-based behavioural measures of deep learning. Deep learning refers to the ability of actively seeking meaning and integrating information in order to understand the material that is taught (Marton & Säljö 1976). Unlike perseverance, there has been little consensus on the behavioural operationalization of deep learning. But just as Dinsmore and Alexander (2012) pointed out: "conceptual definitions should be the lynchpin to creating a measure or measurement of a construct". Based on the definition of deep learning, acts of seeking and utilizing more information could be well linked to the concept. This idea leads to using the *total numbers of different page visits* extracted from log files of two problem solving units. The number of different page visits reflects the ability of looking for information on different websites in order to make a sound judgement, thus serving as a plausible proxy for deep learning skills.

Behavioural measures of deep learning 1: The total number of different page visits extracted from the task of "The Sprained Ankle - Reliable/Trustworthy Site" (PS_u06b).[8] This task shows links to five websites recommended by a friend on how to treat sprained ankle. The respondents are supposed to read through these websites and find the most reliable and trustworthy site. In addition to the one-page content on each webpage, three webpages provide an additional tab for links to read more about the site or the author. These additional webpages provide information such as opinions expressed by individual writers, certified surgeons, or the president of a commercial equipment supplier, which can assist judgement on the credibility and reliability of the sources. The variable has a mean of 4.51 page visits with a standard deviation of 2.39 visits.[9] **Behavioural measures of deep learning 2:** The total number of different page visits extracted from the task of "The Digital Photography Book Order" (PS_u07). This task provides six links to different vendors of or information on digital photography books. Respondents are asked to buy a book for beginners while staying within a budget of 40 USD in time for a friend's birthday in two weeks. Respondents need to find the most suitable website and place an order. Some websites have additional tabs to check availability, shipping costs, etc. These are necessary to make the correct decision on vendor choice. The variable has a mean of 8.19 visits with a standard deviation of 4.14.

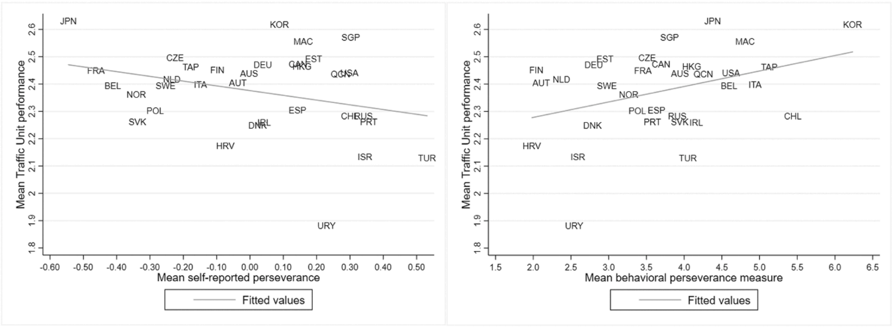
2.2.2.2. Self-reported deep learning. We use the "deep learning" strategy item "looking for additional information" (I_Q04m) as the self-reported measure of deep learning skills. The value ranges from a low level of deep learning: 1 (*not at all*) to a high level: 5 (*to a high extent*). It is a Likert scale measure. But for convenience of comparing with behavioural measures, we treat it as continuous in regressions.[10] The mean level of this variable is 4.05.

[7] Similar as PISA, each respondent in PIAAC only gets a random subset of questions. Therefore, in some countries, certain questions can be missing.

[8] We cannot provide screenshot of PIAAC questions due to confidentiality reasons.

[9] Similar to example one, we show in Section 3.2 that results are insensitive to trimming the number of page visits greater than the 95th percentile.

[10] Table B.2 regresses test scores on self-reported deep learning. Columns (1) and (3) treat deep learning as continuous, while columns (2) and (4) treat it as categorical. Deep learning being one level higher essentially increases score by similar magnitude, indicating that self-reported deep learning can be treated as continuous.



(A) Test performance and self-reported perseverance **(B) Test performance and behavioural perseverance**

Fig. 2. Country average test performance and perseverance.
Note: See Table B.1 in Online Appendix B for the country abbreviations.

Table 2
Perseverance and performance at the individual level.

Variable	(1) Traffic unit performance	(2)	(3)	(4)	(5)
Self-reported perseverance	0.049***		0.028***	0.049***	0.044***
	(0.008)		(0.008)	(0.008)	(0.007)
Behavioural perseverance		0.008***	0.011***	0.008***	0.046***
		(0.001)	(0.001)	(0.001)	(0.007)
Standardised perseverance	No	No	No	No	Yes
Other controls	Yes	Yes	No	Yes	Yes
Observations	14,095	14,152	14,826	14,095	14,095

Note: *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. All the columns report OLS estimates of linear regression models. A constant term is also included in all regressions. ..úOther controls..Ñ include gender, birth year fixed effects, having a computer at home or not, maternal education level, and country fixed effects. In column (5), the self-reported and behavioural perseverance are standardized to mean 0 and standard deviation 1. The number of observations varies across columns due to missing values in perseverance items and control variables.

In Section 3.1, we use a similar data-driven approach as we do for Example 1 to determine the meaning of "the number of different page visits" and construct the corresponding self-reports with PCA for comparison.

2.2.2.3. Outcome variables. **The performance measures for two units: Score for "The Sprained Ankle"** and **Score for "The Book Order"**: Both units have only one correct answer. 1 as correct and 0 as incorrect. The mean score is 0.49 for the "Sprained ankle" unit and 0.48 for the "Book order" unit.

2.2.2.4. Control variable. **Parental education:** We control for the highest level of maternal or paternal education as a proxy for the respondent's innate ability. This categorical variable has 3 levels: 1: Neither parent has attained upper secondary education; 2: At least one parent has attained secondary and post-secondary, non-tertiary education; 3: At least one parent has attained tertiary education. The majority (45.8%) belongs to the second level.

ICT skills: Similar as in Example 1, in case the behavioural measures capture the ability of using computers and the Internet, we also control for respondents' ICT skills at home. It is a continuous index ranged from −1.8 to 6.5, with a higher score indicating more skilful with ICT. The sample mean skill is 2.17, and the standard deviation is 0.91.

We further control for gender (52% of the sample are girls), country fixed effects, and age fixed effects (mean age 37.6, standard deviation 14.1) in the prediction exercises.[11]

2.2.3. Descriptive statistics

Table 4 shows consistency amongst measures of deep learning. The two behavioural measures of deep learning have a positive but modest correlation with the self-reported measure ($\rho_{\text{self,behav1}}$ =0.076 and $\rho_{\text{self,behav2}}$=0.096). And the two behavioural measures have a stronger positive correlation with each other ($\rho_{\text{behav1,behav2}}$=0.558).

Table 3
Perseverance and performance at the country level.

Variable	(1) Average Traffic unit performance	(2)	(3)	(4)	(5)	(6)	(7)
Avg. Self-reported pesev.	−0.173*	−0.065			−0.194**	−0.132	−0.118
	(0.097)	(0.123)			(0.090)	(0.118)	(0.106)
Avg. behavioural persev.			0.057**	0.054*	0.061**	0.063**	0.349**
			(0.025)	(0.028)	(0.024)	(0.028)	(0.159)
Standardised persev.	No	No	No	No	No	No	Yes
Other controls	No	Yes	No	Yes	No	Yes	Yes
Observations	33	32	33	32	33	32	32

Note: All the columns report OLS estimates of linear regression models. *Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. All variables are country averages. A constant term is included in all regressions. "Other controls" include country average probability of being a girl, country average probability of having a computer at home, country mode level of maternal education, and country mode year of birth. In column (7), the average self-reported and behavioural perseverance are standardized to mean 0 and standard deviation 1. The number of observations varies across columns due to missing values in perseverance items and control variables.

Deep learning, as a desirable non-cognitive skill, is supposed to positively correlate with cognitive test performance, both at the

[11] In Section 3.2, we include *immigration status, number of people living in the household, education resources at home*, and *self-reported health* as additional controls. The results are insensitive to further controls. We therefore keep a parsimonious specification for the main analysis.

Table 4
Correlation matrix of test performance and deep learning.

Variable	"Sprain ankle" score	"Book order" score	Self-reported deep learning	Behavioural dp. lrn.1
"Sprain ankle" score	–			
"Book order" score	0.1796***	–		
Self-reported deep learning	0.0330***	0.0637***	–	
Behavioural dp. lrn.1	0.2509***	0.3266***	0.0756***	–
Behavioural dp. lrn.2	0.2501***	0.6554***	0.0961***	0.5575***

Note: * Significant at 10%; ** at 5%; *** at 1%. The correlation between variables in the “Sprain ankle” unit is based on a sample of 20,015 respondents. The sample size for the ..úBook order..Ñ unit is 20,167. Not everyone answers questions from both units because each respondent only gets a random subset of questions. Only about 10,283 individuals receive questions from both units. And 9,983 individuals give answers to both questions.

individual and country levels. At the individual level, Table 4 indeed shows that the self-reported measure of deep learning has a modest positive correlation with scores for two units ($\rho_{\text{self,score1}}$=0.033 and $\rho_{\text{self, score2}}$=0.064), and behavioural measures have a stronger positive correlation with the test performance ($\rho_{\text{behav1,score1}}$=0.251 and $\rho_{\text{behav2, score2}}$=0.655).

However, when aggregated to the country level, self-reported deep learning either displays no correlation (Fig. 3 (A.1)) or a paradoxical negative correlation (Fig. 3 (A.2)) with the test performance, while behavioural measures still consistently show a strong positive correlation (Fig. 3 (B.1) and (B.2)). This may again relate to the group-specific reporting behaviour bias in self-reports. For example, in Fig. 3 (A.1) and (A.2), the Netherlands and the U.S. report themselves to have the lowest and highest level of deep learning. But in Fig. 3 (B.1) and (B.2), the behavioural measures actually show a higher level of deep learning for Dutch respondents. This pattern is similar to the findings in Kapteyn et al. (2007), where they compare the reporting behaviour of work disability between American and Dutch workers. They find that Dutch workers are more “pessimistic” about their health conditions and systematically more likely to report work disability. Here in the deep-learning example, the lowest self-reported deep-learning could be a result of Dutch workers’ conservative reporting style.

This country-level correlation analysis adds support to using behavioural measures, especially when group-specific reporting styles exist.

2.2.4. *Predictive performance on cognitive test scores*

To check if the individual level of positive correlation between deep learning and test performance is driven by individual characteristics, we control for innate ability, ICT skills, demographic characteristics and country fixed effects and use self-reported and behavioural measures of deep learning to predict the performance of “The Sprained Ankle” and “The Book Order” units. Table 5 reports the regression results. In all specifications, self-reported deep learning is no longer predictive to test performance once controlling for individual confounding factors. In contrast, the two behavioural measures are robust to adding further controls, and consistently show a positive association with test performance.

Aggregating variables to the country level, we have only 10 to 13 observations.[12] Table 6 summarises the regression results. Due to the small sample size, in most specifications none of the measures are significant except for columns (3) and (7) for behavioural measures. But comparing the point estimates, the self-reported deep learning is negatively associated with test scores, echoing the paradoxical pattern in Fig. 3 (A.1) and (A.2). While for both behavioural measures, we still observe a positive association with the test performance at the country level.

In sum, our log-based behavioural measures outperform self-reports in predicting test performance at both individual and country levels.

2.3. *Predicting schooling and labour market outcomes*

To show the external validity of our log-based behavioural measures, we further use deep-learning measures to predict schooling and labour market outcomes in PIAAC data, and compare the predictive performance with self-reported measures.[13]

We analyse the association with the following outcome variables: (a) **Years of education:** a continuous variable derived from the highest levels of education. (b) **Training or not:** a dummy indicating if the respondent took non-formal education (for job or non-job-related reasons) in the past year. (c) **Work or not:** a dummy for whether the respondent worked at a job or for any business in the last 12 months. (d) **Manage or not:** a dummy indicating whether the respondent was managing other employees or not. (e) **Skill level of occupation**: a categorical variable describing the skill level of the respondent’s current occupation (the four levels consist of: elementary occupations, semi-skilled blue-collar occupations, semi-skilled white-collar occupations, and skilled occupations). (f) **Hourly earning decile:** a categorical variable of hourly earnings including bonuses for wage and salary earners, in deciles. (g) **Monthly earning decile:** a categorical variable of monthly earnings including bonuses for wage and salary earners and self-employed individuals, in deciles. (h) **Annual income quintile:** a categorical variable of annual income before taxes and deductions, in quintiles. See Section 1 of Online Appendix C for detailed definitions of these outcome variables.

We control for individual confounding factors (e.g., innate ability, ICT skills, demographic characteristics, and country fixed effects) in all specifications.

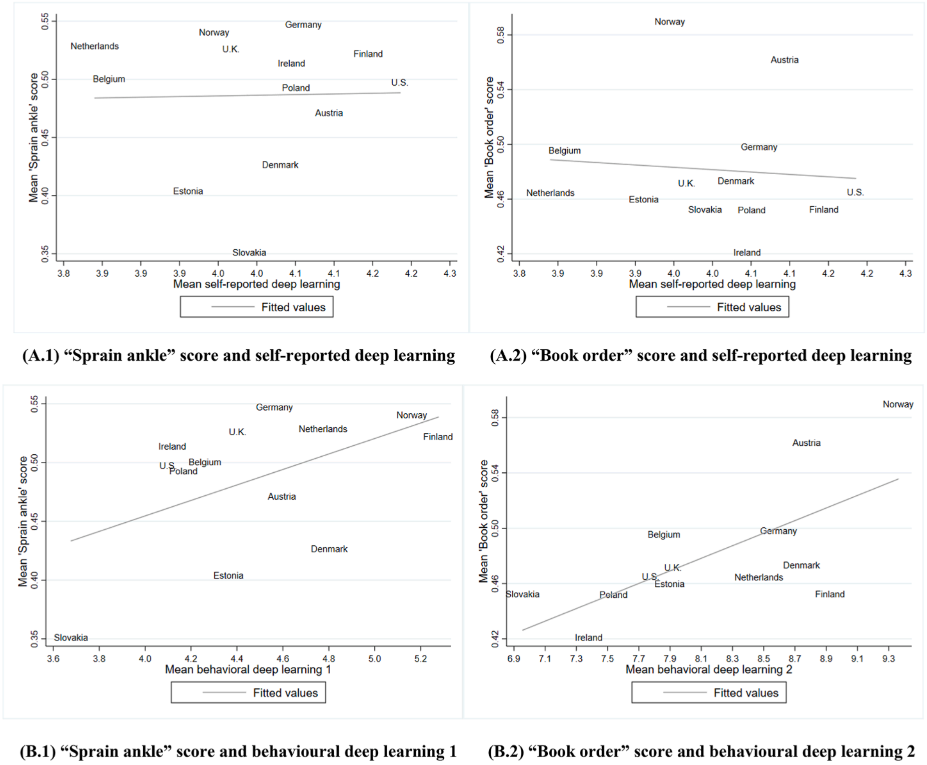
We expect that conditional on individual characteristics, deep learning should still positively influence the schooling and labour market outcomes. Table 7 reports the prediction results.

The self-reported deep learning and two behavioural measures are standardised for the convenience of comparison. Columns (1) and (2) report OLS estimates of linear regression models. Columns (3) to (8) and columns (9) to (16) are estimates of logit models and ordered logit models, respectively. The estimated coefficients of these non-linear models are interpreted as the change in the latent scale (log odds for logit models) of the outcome induced by one standard deviation change in the standardised predictor, ceteris paribus.

In most columns, both self-reported and behavioural measures show significant positive associations with schooling and labour market outcome, except for labour market participation. One standard deviation higher deep learning increases 0.19 to 0.55 years of education, depending on the measure. The higher the level of deep learning, the more likely the respondents are to further their human capital accumulation process by taking training. Self-reported deep learning only

[12] The central limit theorem and the law of large numbers may not hold with this small sample size. So we only use this exercise in Table 6 as a suggestive evidence on the partial correlation between country-level deep learning measures and test scores.

[13] Since PISA test is for 15-year-old students, we do not observe their completed education or labor market outcomes. While PIAAC test, designed for adult workers, collects education and labor market information needed for this exercise. Note that each PIAAC respondent receives only a random subset of questions. The more variables we include in a regression, the fewer individuals who have all variables non-missing in the sample. Therefore, the sample size in this predicting exercise varies largely from 4,467 to 7,320 observations across outcome variables.



(A.1) "Sprain ankle" score and self-reported deep learning

(A.2) "Book order" score and self-reported deep learning

(B.1) "Sprain ankle" score and behavioural deep learning 1

(B.2) "Book order" score and behavioural deep learning 2

Fig. 3. Country average test performance and deep learning.

Table 5
Deep learning and performance at the individual level.

Variable	(1) "Sprain ankle" score	(2)	(3)	(4)	(5) "Book order" score	(6)	(7)	(8)
Self-reported deep learning	−0.005 (0.005)		−0.008 (0.005)	−0.006 (0.004)	0.007 (0.005)		−0.003 (0.004)	−0.003 (0.003)
Behavioural dp. lrn. 1		0.045*** (0.002)	0.045*** (0.002)	0.108*** (0.004)				
Behavioural dp. lrn. 2						0.079*** (0.001)	0.079*** (0.001)	0.328*** (0.003)
Standardised dp. lrn.	No	No	No	Yes	No	No	No	Yes
Other controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	14,698	14,680	14,679	14,679	14,708	14,692	14,688	14,688

Note: * Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. All the columns report OLS estimates of linear regression models. A constant term is also included in all regressions. "Other controls" include gender, age fixed effects, ICT skills, parental education levels, and country fixed effects. In columns (4) and (8), self-reported and behavioural measures of deep learning are standardized to mean 0 and standard deviation 1. The number of observations varies across columns due to missing values in deep learning variables and control variables.

Table 6
Deep learning and performance at the country level.

Variable	(1) Avg. "Sprain Ankle" score	(2)	(3)	(4)	(5) Avg. "Book Order" score	(6)	(7)	(8)
Avg. Self-reported dp. lrn.	−0.031 (0.234)		0.011 (0.137)	−0.149 (0.201)	−0.152 (0.152)		−0.046 (0.091)	−0.191 (0.143)
Avg. Behavioural dp. lrn.1		0.091 (0.053)	0.066* (0.036)	0.105 (0.060)				
Avg. Behavioural dp. lrn.2						0.032 (0.033)	0.046** (0.015)	0.040 (0.031)
Standardised dp. lrn.	No	No	No	No	No	No	No	No
Other controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Observations	10	10	13	10	10	10	13	10

Note: All the columns report OLS estimates of linear regression models. * Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. All variables are country averages. A constant term is included in all regressions. "Other controls" include country average probability of being a girl, country average ICT skills, country mode level of parental education, and country average age. The number of observations varies across columns due to missing values in perseverance items and control variables.

Table 7
Predicting schooling and labour market outcomes with deep learning.

Variable	Years of education		Training or not		Work or not		Manage or not	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Self-reported deep learning	0.190***	0.194***	0.077***	0.115***	0.024	0.043*	0.146***	0.140***
	(0.027)	(0.027)	(0.020)	(0.020)	(0.024)	(0.024)	(0.026)	(0.026)
Behavioural dp. lrn.1	0.465***		0.229***		0.101***		0.077***	
	(0.027)		(0.019)		(0.024)		(0.025)	
Behavioural dp. lrn.2		0.549***		0.220***		0.154***		0.115***
		(0.027)		(0.019)		(0.025)		(0.025)
Observations	13,445	13,441	13,626	13,659	14,681	14,691	9509	9456

Variable	Skill level of occupation		Hourly earning decile		Monthly earning decile		Annual income quintile	
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Self-reported deep learning	0.141***	0.135***	0.054***	0.090***	0.070***	0.107***	0.057***	0.100***
	(0.019)	(0.019)	(0.020)	(0.021)	(0.019)	(0.020)	(0.020)	(0.020)
Behavioural dp. lrn.1	0.265***		0.239***		0.220***		0.215***	
	(0.018)		(0.020)		(0.019)		(0.020)	
Behavioural dp. lrn.2		0.311***		0.281***		0.259***		0.253***
		(0.019)		(0.020)		(0.019)		(0.020)
Observations	12,523	12,522	8922	8937	9782	9810	9770	9799

Note: * Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. In all columns, the self-reported and behavioural measures of deep learning are standardized to mean 0 and standard deviation 1. Gender, age fixed effects, ICT skills, parental education levels, and country fixed effects are included. Columns (1) and (2) report OLS estimates of linear regression models. Columns (3) to (8) report estimates of logit models. Columns (9) to (16) report estimates of ordered logit models. The coefficients of non-linear models are changes in the latent scales (log odds for logit models) of the outcome induced by one standard deviation change in the standardized predictor, ceteris paribus. A constant term is included in columns (1) to (8). The number of observations varies across columns due to random missing values in deep learning variables, outcomes, and control variables.

shows an insignificant and weak positive relationship with labour market participation, while the behavioural measure suggests that individuals with higher deep learning levels are significantly more likely to work. And a higher level of deep learning, regardless of the measure, is associated with a higher probability of being in a management position, a higher level of skills used at work, and a higher level of earnings.

Comparing the standardised measures of deep learning, behavioural measures predict a comparable, and sometimes larger effect on schooling and labour market outcomes than self-reports. But the fact that all three measures show predictive validity in schooling and labour market outcomes indicates the complementarity of self-reported and behavioural measures at least at the individual level.

3. Sensitivity analysis

3.1. A data-driven method to interpret log-based behavioural measures and construct self-reported assessments

In Section 2, we construct log-based behavioural measures and self-reports based on the definition of the desired non-cognitive traits, mainly to showcase the predictive performance of the log-based measure. This conceptual definition-based interpretation of behavioural measures is essentially a top-down approach. We now check how sensitive our results will be if we experiment with a data-driven method for interpretation and comparison. First, we investigate how log-based behavioural measures correlate with various possibly related self-reported traits. The most relevant traits might help shed light on what behavioural measures are about. Second, these most relevant traits are used to construct corresponding self-reported measures. Finally, we repeat predicting exercises in Section 2 with log-based measures and these data-driven self-reports. We show that the data-driven interpretation of behavioural measures has some consistency with our main interpretation, and that the predictive advantage of log behavioural measures remains.

3.1.1. Interpret the behavioural measures and construct the self-reported measures

We check if “total number of resets” is also related to the following self-reported traits besides perseverance: (a) **Perceived control:** a continuous measure extracted with factor analysis from 6 perceived control-related items about math learning. (b) **Attitudes toward school:** An index of students’ attitudes toward school learning outcomes constructed by PISA. (c) **Open to problem solving:** a PISA-constructed index of openness to problem solving. (d) **Math anxiety:** a PISA-constructed index of anxiety about math. (e) **Cognitive activation:** a PISA-constructed index of how often the respondent’s teacher uses cognitive activation strategies in math class. (f) **Instrumental motivation:** a PISA-constructed index of instrumental motivation to learn mathematics.[14]

For “number of page visits” in PIAAC, besides deep learning strategy “looking for additional information”, we check if it is related to other learning strategies: “relate new ideas into real life”, “like learning new things”, “attribute something new”, “get to the bottom of difficult things”, and “figure out how different ideas fit together”. They are all Likert scale measures (1: “not at all” to 5: “to a high extent”). We also check if “number of page visits” is related to “trust only few people” and “careful about being taken advantage of”. These two Likert scale measures have values from 1: “strongly agree” to 5: “strongly disagree”. A larger value can be seen as less sceptical or more trusting/open. See Online Appendix C for detailed definitions of these non-cognitive traits and Table B.3 of Online Appendix B for summary statistics.

Table 8 summarises the Pearson correlation and partial correlation between behavioural measures and self-reported traits. In columns (1) and (2), “total number of resets” is positively correlated with “perseverance” and “attitudes toward school”. The correlation is statistically significant (though modest) and robust to controlling for individual confounders.[15] Students with more reset clicks tend to be more perseverant and strongly motivated for a better schooling outcome. Note that “perseverance” and “being motivated for goals” are key components of Duckworth et al.’s (2007) definition of grit: “perseverance and passion for long-term goals”. This suggests that the behaviour of reset clicks indeed captures grit-related traits, showing some consistency with our

[14] For details of PISA-constructed variables, please refer to the “PISA 2012 Technical Background” at https://doi.org/10.1787/9789264208094-10-en.

[15] It is not uncommon to see correlation < 0.1 between task-specific and non task-specific measures, e.g., task-specific behavioural measures vs. general self-reports (Van Hout-Wolters 2009, Zamarro et al. 2018). Yet given the two measuring methods compared, a robust and statistically significant correlation still suggests some convergent validity of the two measures.

Table 8
Correlation between behavioural measures and self-reported traits.

	Total number of resets			#Page visits in “Sprain ankle”		#Page visits in “Book order”	
	Corr.	Partial corr.		Corr.	Partial corr.	Corr.	Partial corr.
Self-reports:	(1)	(2)	Self-reports:	(3)	(4)	(5)	(6)
Perseverance	0.017**	0.014*	Relate new ideas into real life	0.133***	0.056***	0.133***	0.063***
Perceived control	0.009	0.011	Like learning new things	0.126***	0.038***	0.130***	0.028***
Attitudes toward school	0.025**	0.028**	Attribute something new	0.121***	0.049***	0.128***	0.053***
Open to problem solving	−0.022***	−0.008	Get to the bottom of difficult things	0.070***	0.027***	0.078***	0.033***
Math anxiety	0.015	−0.011	Figure out how different ideas fit together	0.047***	−0.002	0.060***	0.016*
Cognitive activation	−0.019*	−0.011	Looking for additional info.	0.076***	0.018**	0.096***	0.025***
Instrumental motivation	−0.006	0.003	Trust only few people	0.134***	0.090***	0.132***	0.084***
			Careful about being taken advantage of	0.164***	0.105***	0.160***	0.106***

Note: * Significant at 10%; ** at 5%; *** at 1%. The self-reported non-cognitive traits and behavioural measures in all columns are standardized to mean zero and standard deviation one. Odd columns report Pearson correlations and even columns report partial correlations from regressions. Column (2) controls for gender, birth year fixed effects, having a computer at home or not, maternal education level, and country fixed effects. Columns (4) and (6) include gender, age fixed effects, ICT skills, parental education levels, and country fixed effects. The number of observations varies across grids due to random missing values in self-reports, behavioural measures, and control variables.

top-down interpretation in Section 2.

“Total number of page visits” in columns (3) to (6) is modestly but robustly correlated with most dimensions of learning strategies. As noted by McNamara (2011), there can be shared variance between underlying deep learning skills and uses of learning strategies, and the shared variance may stem from the similarities across measures of learning strategies. Therefore, if the number of page visits indeed captures deep learning skills, we will not be surprised to see it correlated with a number of interrelated learning strategies. Columns (3) to (6) also show that respondents with more page visits tend to be less sceptical and more trusting or open. Taken together, the behaviour of page visits is likely to reflect some deep learning-related skills and some openness to more information, which is not far from our interpretation in Section 2.

We then select the most relevant self-reported traits, i.e., those with robust and statistically significant Pearson and partial correlations, to construct the one-dimensional self-reports with PCA. In PISA, **self-reported measure** for “total number of resets” is extracted from “perseverance” and “attitudes toward school”, with a mean of 0 and a standard deviation of 1.12.[16] In PIAAC, **self-reported measure 1** for “total page visits in ‘Sprain Ankle’ Unit” is extracted from all traits in column (4) of Table 8 except for “Figure out how different ideas fit together”. The measure has a mean of 0 and a standard deviation of 1.65. **Self-reported measure 2** for “total page visits in ‘Book order’ Unit” is extracted from all traits in column (6), with a mean of 0 and a standard deviation of 1.80. Table B.4 in Online Appendix B reports the scoring coefficients for all the selected traits. The PCA-constructed self-reports in PIAAC load heavily on items of learning strategies and less so on items about trusting or openness.

3.1.2. *Predictive performance revisited*

We revisit the main predicting exercises in Section 2 to check whether log-based behavioural measures still outperform PCA-constructed self-reports.

Fig. 4 plots the country average test scores and constructed self-reported measures in PISA and PIAAC. Note that plots for scores and behavioural measures are unchanged, as seen in Fig. 2(B), 3(B.1), and 3 (B.2). Fig. 4(A), 4(B), and 4(C) replicate the pattern of Fig. 2(A), 3(A.1), and 3(A.2), displaying a paradoxical negative relationship. This implies that the constructed self-reports, which correspond to the data-driven interpretations of behavioural measures, still fail to retain an intuitive positive correlation with test performance at the country level.

Table 9 compares how behavioural and constructed self-reported measures predict the test scores controlling for individual characteristics. At the individual level, columns (1), (3), and (4) are very close to column (5) of Table 2 and columns (4) and (8) of Table 5. Both self-reported and behavioural measures in PISA predict test performance well, while self-reported measures in PIAAC no longer predict test performance once controlling for individual confounding factors. At the country level, columns (2), (5), and (6) essentially replicate the pattern of column (7) of Table 3 and columns (4) and (8) of Table 6. The partial correlations between country average self-reports and test scores are all negative, echoing the pattern in Fig. 4. These results imply a better predictive validity of log-based behavioural measures against constructed self-reports, especially at the country level.

Similar to Tables 7 and 10 shows that both log-based behavioural measures and constructed self-reports predict a comparable effect on schooling and labour market outcomes at the individual level.

In summary, with the data-driven method, the interpretation of behavioural measures has shown some consistency with the conceptual definition-based interpretation. And the essential feature that log-based behavioural measures have higher predictive validity than self-reports at both individual and country levels still remains.

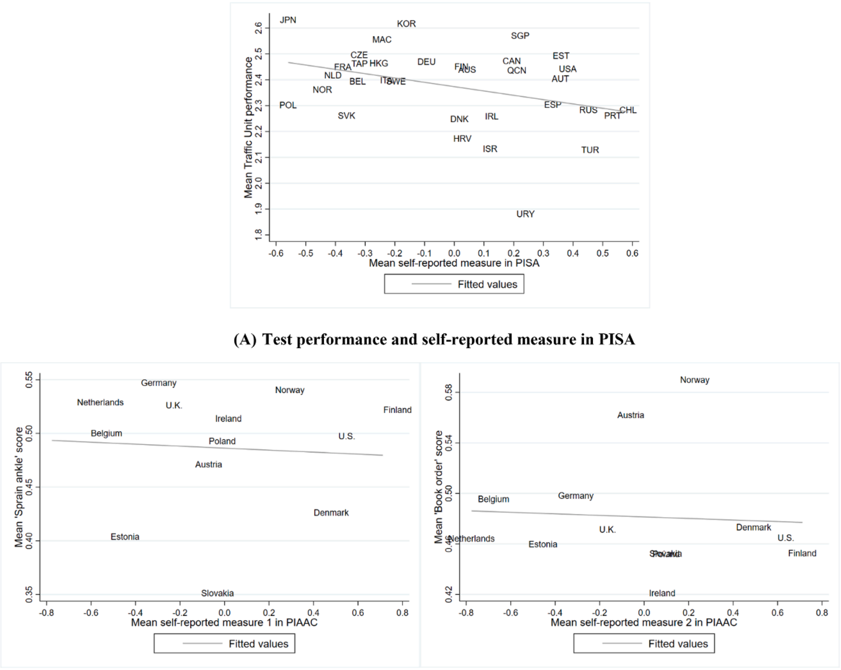
3.2. *Trimmed behavioural measures and additional controls*

We check if the main results are sensitive to the distribution of log-based measures and additional controls. We trim the “total number of resets” and the two “number of page visits” for values larger than the 95th percentile, in case that too many resets or page visits would capture noise (e.g., stubbornness or meaningless random clicks) rather than the desired non-cognitive traits.

To further account for potential individual confounding factors, we include the following variables in PISA: *immigration status* (a categorical variable of immigration background), *educational resources at home* (an index for how much educational resources the student has at home), *wealth* (an index of family wealth), *whether mom/dad/brother/sister at home* (four dummy variables for mom/dad/brother/sister usually living at home with the respondent), and *friends' performance in math* (a categorical variable of how well the student's friends do in math). The *friends' performance in math* can be seen as a proxy for the respondent's own cognitive ability as well as peer influence. For PIAAC regressions, we include *immigration status, number of books at home* (a categorical variable of how many books at home), *health* (a categorical variable of self-reported health with values from 1 “Poor” to 5 “Excellent”), and *number of people living in the household* (top coded at 6). See Online Appendix C for detailed definitions of these controls.

Table B.5 in Online Appendix B predicts test scores with trimmed behavioural measures and additional controls at both individual and country levels. And Table B.6 predicts schooling and labour market

[16] Only 7383 observations have values for this self-reported measure because only about half the sample (7414 observations) got the question about “attitudes toward learning”.



(A) Test performance and self-reported measure in PISA

(B) "Sprain ankle" score and self-reported measure 1 in PIAAC **(C) "Book order" score and self-reported measure 2 in PIAAC**

Fig. 4. Country average test performance and PCA-constructed self-reported measures.

Table 9
Predicting test performance with behavioural and constructed self-reports.

Variable	PISA-Indiv.	PISA-Cnt.	PIAAC-Indiv.		PIAAC-Cnt.	
	Score (1)	Avg. score (2)	Score 1 (3)	Score 2 (4)	Avg. score 1 (5)	Avg. score 2 (6)
(Avg.) Self-reported measure	0.052*** (0.010)	−0.089 (0.111)				
(Avg.) Behavioural measure	0.053*** (0.010)	0.308* (0.156)				
(Avg.) Self-reported measure 1			0.006 (0.005)		−0.066 (0.092)	
(Avg.) Behavioural measure 1			0.107*** (0.004)		0.258 (0.147)	
(Avg.) Self-reported measure 2				0.001 (0.004)		−0.059 (0.068)
(Avg.) Behavioural measure 2				0.328*** (0.003)		0.159 (0.145)
Observations	7035	32	14,630	14,620	10	10

Note: * Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. The behavioural and constructed self-reported measures in all columns are standardized to mean zero and standard deviation one. Column (1) regresses the score of ..úTraffic unit..Ñ on the behavioural measure (total number of resets) and the PCA-constructed self-reported measure in Section 3.1.1, using PISA data at the individual level. Gender, birth year fixed effects, having a computer at home, maternal education level, and country fixed effects are also included. Column (2) performs a similar regression at the country level. All variables are country averages. And the country average probability of being a girl, country average probability of having a computer at home, country mode level of maternal education, and country mode year of birth are included. Column (3) regresses the "Sprain ankle" score (Score 1) on behavioural measure 1 (#page visits in the "Sprain ankle" unit) and PCA-constructed self-reported measure 1 in Section 3.1.1. Column (5) regresses the "Book order" score (Score 2) on behavioural measure 2 (#page visits in "Book order" unit) and PCA-constructed self-reported measure 2. Both columns use data from PIAAC at the individual level. Gender, age fixed effects, ICT skills, parental education levels, and country fixed effects are also included. Columns (4) and (6) perform similar regressions but at the country level. All variables are country averages. We also control for country average probability of being a girl, country average ICT skills, country mode level of parental education, and country average age. The number of observations varies across columns due to random missing values in self-reports, behavioural measures, and control variables.

outcomes, similarly to Table 7. The results are very close to those in Section 2.

4. Discussion

We set out to introduce an innovative source of behavioural measures, using log-file data to construct behavioural measures. Log-based behavioural measures are unobtrusive to collect and immune to self-presentation styles and reference group effects. We compare them with self-reported measures for predictive performance in cross-cultural contexts, using existing data from the large-scale PISA and PIAAC studies. We show that log-based behavioural measures have better predictive performance than self-assessments, as they consistently have power to predict the test performance both at the individual and the country level, whereas self-reported measures produce weaker, and sometimes counter-intuitive associations, especially in cross-group (country) comparisons. The log-based behavioural measures also show predictive validity in schooling and labour market outcomes, showing

Table 10
Predicting schooling and labour market outcomes with behavioural and constructed self-reports.

Variable	Years of education		Training or not		Work or not		Manage or not	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Self-reported measure 1	0.499***		0.240***		0.193***		0.318***	
	(0.029)		(0.021)		(0.026)		(0.028)	
Behavioural measure 1	0.437***		0.218***		0.087***		0.066***	
	(0.027)		(0.019)		(0.024)		(0.025)	
Self-reported measure 2		0.431***		0.246***		0.186***		0.283***
		(0.029)		(0.021)		(0.026)		(0.028)
Behavioural measure 2		0.529***		0.213***		0.143***		0.104***
		(0.027)		(0.020)		(0.025)		(0.025)
Observations	13,399	13,376	13,580	13,594	14,632	14,623	9480	9413

Variable	Skill level of occupation		Hourly earning decile		Monthly earning decile		Annual income quintile	
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Self-reported measure 1	0.329***		0.186***		0.199***		0.187***	
	(0.021)		(0.022)		(0.021)		(0.022)	
Behavioural measure 1	0.251***		0.228***		0.210***		0.206***	
	(0.019)		(0.020)		(0.019)		(0.020)	
Self-reported measure 2		0.310***		0.182***		0.203***		0.201***
		(0.020)		(0.022)		(0.021)		(0.021)
Behavioural measure 2		0.299***		0.269***		0.248***		0.242***
		(0.019)		(0.020)		(0.019)		(0.020)
Observations	12,483	12,467	8892	8898	9750	9767	9738	9756

Note: * Significant at 10%; ** at 5%; *** at 1%. Numbers in parentheses are standard errors. In all columns, the behavioural and constructed self-reported measures are standardized to mean zero and standard deviation one. Gender, age fixed effects, ICT skills, parental education levels, and country fixed effects are included. Behavioural measures 1 and refers to #page visits in "Sprain ankle" and "Book order" units, respectively. Self-reported measures 1 and 2 are the PCA-constructed measure with PIAAC data in Section 3.1.1. Columns (1) and (2) report OLS estimates of linear regression models. Columns (3) to (8) report estimates of logit models. Columns (9) to (16) report estimates of ordered logit models. The coefficients of non-linear models are changes in the latent scales (log odds for logit models) of the outcome induced by one standard deviation change in the standardized predictor, ceteris paribus. A constant term is included in columns (1) to (8). The number of observations varies across columns due to random missing values in predictors, outcomes, and control variables.

promise for a wide range of applications.

4.1. *Complementing self-reports with unobtrusive behavioural measures*

Log-based behavioural measures have clear advantages over self-reports and lab-based performance tasks. First, they are less sensitive to self-presentation styles and thus more objective; second, they occur in a natural environment instead of in contrived lab-settings and are thus unobtrusive, and third, they are easy to implement (with well-developed and validated tasks) in computer-based assessment, allowing to reach a number of respondents significantly larger than could be achieved in lab settings. The log-based behavioural measures will be especially useful in cross-group/country studies, particularly for data where vignette questions to correct for differences in subjective response scales used in self-assessments are not available. Log-based behavioural measures can also be used to cross-check the validity of the self-reported measures.

Still, log-based behavioural measures are limited by their availability and specificity. First, they are, obviously, only available when log-file data are available. Given the progress of online data collection methods and the increasing number of newly released datasets, we expect to see more log-file data becoming available in the future. Second, the log-based behavioural measure is restricted by the specificity of the task and usually only speaks to one or a limited number of facets of non-cognitive skills, which may limit the generalization of the empirical findings. This limitation of specificity also applies to other behavioural measures (including observational behaviours and lab-based performance task measures). To investigate multi-dimensional non-cognitive skills, researchers will have to extract multiple log-based behavioural measures from different domains of tasks, and use dimension reduction techniques if necessary. Self-reported measures, on the other hand, can easily capture the multi-dimensional nature of non-cognitive skills, either by asking respondents to give an overall rating of non-cognitive skills, or by applying dimension reduction techniques (e.g., principal component analysis) to construct a measure from multiple Likert-scale items. It therefore makes sense to use both log-based behavioural measures and self-reported measures complementarily.

4.2. *Future directions*

Our study is a first step to investigate the potential of this type of measures. Limited by the cross-sectional nature of the data, we can only check the cross-sectional predictive validity in test performance, schooling, and labour market outcomes. A natural next step would be to investigate the longitudinal predictive validity of log-based behavioural measures and education or labour market outcomes (e.g., linking PISA data to education and labour market related administrative data to investigate the long-run effects of non-cognitive skills). As for the interpretation of log-based behavioural measures, the conceptual definition-based and data-driven methods discussed in this study are rather exploratory. Possible alternative interpretations of these behavioural measures are yet to be explored. Future research should continue investigating the validity of these interpretations across a variety of circumstances. And much more work in the future is needed to investigate the psychometric properties of log-based behavioural measures. Furthermore, all of our behavioural measures are one-dimensional indicators from the domain of problem-solving. Future studies can design and validate a large variety of tasks in different domains (e.g., reading, numeracy, problem solving) and extract multiple indicators. Dimension reduction techniques can then be considered for constructing a multi-faceted log-based behavioural measure.

Data Availability

The authors do not have permission to share data.

Acknowledgement

We thank the PIAAC team and GESIS – Leibniz Institute for the Social Sciences for providing the PIAAC log data, and Judith Pal for pre-processing the PISA log file. This work was supported by the National Natural Science Foundation of China (grant number: 72273163), and by the Joint Usage and Research Center, Institute of Economic Research, Hitotsubashi University (grant ID: IERPK2213). We are very grateful for many helpful comments from the editor and the anonymous referee. We thank Jochem de Bresser, Bianca De Stavola, Isaac Ehrlich, Colm Harmon, Marike Knoef, Jan van Ours, Arthur van Soest, and seminar participants at IAIR 2019 Shanghai Conference and CIES 2019 San Francisco for their helpful comments. There are no conflicts of interest.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.socec.2023.101992.

References

Bond, T. N., & Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy, 127*(4), 1629–1640.

Chen, Y., Feng, S., Heckman, J. J., & Kautz, T. (2020). Sensitivity of self-reported noncognitive skills to survey administration conditions. *Proceedings of the National Academy of Sciences, 117*(2), 931–935.

Dinsmore, D. L., & Alexander, P. A. (2012). A critical discussion of deep and surface processing: What it means, how it is measured, the role of context, and model specification. *Educational psychology review, 24*(4), 499–567.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*, 1087. https://doi.org/10.1037/0022-3514.92.6.1087

Eisenberger, R. (1992). Learned industriousness. *Psychological Review, 99*(2), 248–267.

Feng, S., Han, Y., Heckman, J. J., & Kautz, T. (2022). Comparing the reliability and predictive power of child, teacher, and guardian reports of noncognitive skills. *Proceedings of the National Academy of Sciences, 119*(6), Article e2113992119.

Garcia, E. (2016). The need to address non-cognitive skills in the education policy agenda. *Non-cognitive skills and factors in educational attainment* (pp. 31–64). Brill Sense.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics, 24*(3), 411–482.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review, 91*(2), 145–149.

Hitt, Collin, Trivitt, Julie, & Cheng, Albert (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review, 52*, 105–119.

Kapteyn, A., Smith, J. P., & Van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review, 97*(1), 461–473.

Konstabel, K., Aavik, T., & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality, 20*, 549–566. https://doi.org/10.1002/per.593

Kyllonen, P. C., & Bertling, J. J. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–286). Boca Raton, FL: CRC Press.

Leising, D., Locke, K. D., Kurzius, E., & Zimmermann, J. (2015). Quantifying the association of self-enhancement bias with self-ratings of personality and life satisfaction. *Assessment, 23*, 588–602. https://doi.org/10.1177/1073191115590852

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*(1), 101–128.

Lundberg, S. (2015). *Non-cognitive skills as human capital.* Santa Barbara: University of California.

Markman, G. D., Baron, R. A., & Balkin, D. B. (2005). Are perseverance and self-efficacy costless? Assessing entrepreneurs' regretful thinking. *Journal of Organizational Behavior, 26*(1), 1–19.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British journal of educational psychology, 46*(1), 4–11.

McNamara, D. S. (2011). Measuring deep, reflective comprehension and learning strategies: Challenges and successes. *Metacognition and Learning, 6*(2), 195–203.

Määttänen, I., Makkonen, E., Jokela, M., Närväinen, J., Väliaho, J., Seppälä, V., et al. (2021). Evidence for a Behaviourally Measurable Perseverance Trait in Humans. *Behavioral sciences (Basel, Switzerland), 11*(9), 123. https://doi.org/10.3390/bs11090123

OECD. (2013a). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD Publishing.

OECD. (2013b). *PISA 2012 technical report.* Paris, France: OECD Publishing.

OECD. (2017). Programme for the International Assessment of Adult Competencies (PIAAC) log files (Publication no. 10.4232/1.12955). (ZA6712 Data file Version 2.0.0). from GESIS Data Archive.

Renninger, K. A., & Bachrach, J. E. (2015). Studying triggers for interest and engagement using observational methods. *Educational Psychologist, 50*, 58–69. https://doi.org/10.1080/00461520.2014.999920

Reynolds, B., Ortengren, A., Richards, J. B., & de Wit, H. (2006). Dimensions of impulsive behavior: Personality and behavioral measures. *Personality and Individual Differences, 40*, 305–315. https://doi.org/10.1016/j.paid.2005.03.024

Robert, A. A., Donnellan, M. B., Brent, W. R., & Fraley, R. C. (2015). The effect of response format on the psychometric properties of the Narcissistic Personality Inventory: Consequences for item meaning and factor structure. *Assessment, 23*, 203–220. https://doi.org/10.1177/1073191114568113

Van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology, 43*, 1205–1228. https://doi.org/10.1177/0022022111428083

Van Hout-Wolters, B. H. A. M. (2009). Leerstrategieën meten: Soorten meetmethoden en hun bruikbaarheid in onderwijs en onderzoek. *Pedagogische Studiën, 86*.

Voňková, Hana, & Hullegie, Patrick (2011). Is the anchoring vignette method sensitive to the domain and choice of the vignette? *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(3), 597–620.

Weiss, A. (1988). High school graduation, performance, and wages. *Journal of Political Economy, 96*(4), 785–820.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., et al. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis, 38*, 148–170. https://doi.org/10.3102/0162373715597298

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