Studies in Educational Evaluation 66 (2020) 100880

Contents lists available at ScienceDirect

Studies in Educational Evaluation

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

The role of engagement and academic behavioral skills on young students' academic performance—A validation across four countries

Fazilat Siddiq[a,b,*], Perman Gochyyev[c], Ona Valls[d]

[a] The Nordic Institute for Studies in Innovation, Research and Education (NIFU), Postboks 2815 Tøyen, 0608, Oslo, Norway
[b] University of South-Eastern Norway, Department of Education and Quality in Learning (USN eDU), Norway
[c] Berkeley Evaluation and Assessment Research (BEAR) Center, University of California, Berkeley, USA
[d] Department of Sociology, Autonomous University of Barcelona (UAB), Spain

ARTICLE INFO

Keywords:
Academic behavioral skills
Engagement
Secondary education
Item response theory
Rasch modelling
Validation

ABSTRACT

The aim of this study is to validate an instrument measuring students' academic behavioral skills and engagement—skills identified as vital for student achievement. We inspect the reliability and validity of the survey with respect to item fit, factorial structure, relations with academic performance, and the fairness of the items across student groups. The fairness analyses are critical to making valid comparisons between groups and across countries. Data comprising 8520 grade 10 students from four countries were analysed using item response theory. We found that both scales were multidimensional, acted fairly across students' gender, country, immigrant-, and socio-economic background (after removing four items), and were positively and significantly correlated with self-reported and performance-based academic performance.

1. Introduction

In educational research, there has been an increased focus on concepts such as deeper learning, higher order thinking, self-regulated learning, meta-cognitive learning, non-cognitive skills, and 21st century skills (Farrington et al., 2012; Palardy & Rumberger, 2019; Pellegrino & Hilton, 2012). The enterprises (e.g., educational policymakers, educators and researchers) supporting such concepts share a common understanding that there is a need for a renewed focus on the development of broad and transferable skills and knowledge. Although the various concepts represent different traditions and include different competencies and skills, they share common principles. First, they describe attitudes, perceptions and characteristics as opposed to only cognitive competencies (i.e., academic achievement, intelligence). Second, the role of such competencies and skills have been identified as critical for students' learning and achievement (e.g., Gutman & Schoon, 2013; Pellegrino & Hilton, 2012). Particularly, students' *academic behavioral skills* (e.g., self-efficacy, collaboration, communication, and self-management) and *engagement* have been identified as correlational with academic achievement (Farrington, 2013; Palardy & Rumberger, 2019; Te Wang, Fredricks, Ye, Hofkens, & Linn, 2019).

Moreover, academic behavioral skills and engagement are considered malleable, meaning they can be taught, developed and transformed (Gutman & Schoon, 2013). However, researchers have emphasized that the available research in this field is limited, primarily focusing on cognitive competencies than for example, non-cognitive competencies and skills (e.g., self-concept, social skills, creativity, self-control), and few studies investigated several constructs (Pellegrino & Hilton, 2012). To our knowledge, cross-national comparative studies, including the measurement of several domains in one study, are at infancy. More importantly, validation studies of instruments measuring such skillsets that are used to compare groups (e.g., countries, gender) and utilized to measure school systems and their success are scarce.

Against this background, the current study is aimed at inspecting secondary school students' academic behavioral skills and engagement, and their relations with academic performance in four European countries. Using the International Study of City Youth (ISCY) sample, we investigate (1) the internal validity of the measures; (2) the fairness of the measures across students' gender, national-, immigrant- and socio-economic background; and (3) the relations between the students' academic behavioral skills and engagement, and academic performance.

Researchers accentuate that it is important to assess the validity of adolescent responses in self-report surveys (Clark & Malecki, 2019). This is particularly meaningful in the current study as the included scales are used in an international comparative study. Hence, the overarching aim of this study is to gather and present evidence on the quality of the survey instrument with reference to commonly agreed

* Corresponding author at: The Nordic Institute for Studies in Innovation, Research and Education (NIFU), Postboks 2815 Tøyen, 0608, Oslo, Norway.
E-mail addresses: fazilat.siddiq@nifu.no (F. Siddiq), perman@berkeley.edu (P. Gochyyev), ona.valls@uab.cat (O. Valls).

https://doi.org/10.1016/j.stueduc.2020.100880
Received 28 October 2019; Received in revised form 30 March 2020; Accepted 31 March 2020
Available online 25 June 2020
0191-491X/

aspects of reliability and validity (AERA, APA, & NCME, 2014; Messick, 1995).

2. Theoretical framework

Broad terms such as 21st century skills (Griffin, McGaw, & Care, 2012), non-cognitive skills (Khine & Areepattamannil, 2016), social and emotional skills (OECD, 2015), academic mindsets (Farrington, 2013), deeper learning (Farrington, 2013), grit (Clark & Malecki, 2019; Duckworth, Peterson, Matthews, & Kelly, 2007), and the Big-Five (John & Srivastava, 1999) have been used to describe personality traits, behaviors, skills, attitudes and competencies, which are considered important in education to increase students' learning and achievement (Farrington et al., 2012; Griffin et al., 2012; OECD, 2015; Palardy & Rumberger, 2019; Te Wang et al., 2019). Even though the broad spectra of terms were introduced at different times and include different sets of competencies, they are overlapping (Credé, Tynan, & Harms, 2017), and share the common view that non-cognitive skills as described in these frameworks (e.g., grit, conscientiousness, social skills, self-concept) are equally important for academic achievement as the mere cognitive abilities (Gutman & Schoon, 2013). The initiatives supporting these concepts share the mutual understanding that there is a need for focusing on the development of such broad and transferable skillsets and knowledge (Pellegrino & Hilton, 2012).

As mentioned, there are different theoretical perspectives that include different kinds of non-cognitive skills or inter- and intrapersonal skills. From the research on engagement, we find scholars who incorporate variables such as self-efficacy (Jimerson, Campos, & Green, 2003) or self-regulation (Fredricks, Blumenfeld, & Paris, 2004; Janosz, Archambault, Morizot, & Pagani, 2008; Appleton et al., 2008) within the construction of cognitive engagement. In contrast, authors such as Renninger and Hidi (2016) state that both engagement and motivational variables (goals, self-efficacy and self-regulation) describe the way in which an individual interacts with the environment, in this case, the school.

Another perspective is the 21st century skills framework that considers among others critical thinking, problem-solving, creativity, communication and collaboration as sub-competence areas (Partnership for 21st Century Skills (P21) & AACTE (American Association of Colleges of Teacher Education), 2010, p. 2; Griffin & Care, 2015). Pellegrino and Hilton (2012) classified the 21 st century skills as covering three larger competency clusters: the cognitive domain, the intrapersonal domain and the interpersonal domain. Moreover, they examined the importance of various types of 21st century skills for success in various aspects of life, such as education, work, and health. Some of the key findings were that existing research is limited and primarily correlational. In addition, they argued that the intrapersonal and interpersonal domains hadbeen poorly studied compared to the cognitive competences. Moreover, among the intrapersonal- and interpersonal skills, aspects related to *conscientiousness,* such as staying organized or being responsible and hardworking, are the most highly correlated with positive educational and career outcomes (Pellegrino & Hilton, 2012). Also, Palardy and Rumberger (2019) emphasize the relations between certain dispositions and non-cognitive skills, e.g., self-efficacy and self-control (intra-personal), and social skills such as communication and collaboration (inter-personal). According to these scholars, the 21 st century skills framework includes student dispositions, types of school engagement, intrapersonal- and interpersonal skills (Palardy & Rumberger, 2019). However, there are few studies that have investigated competencies within several domains, and to our knowledge, cross-national comparative studies, including the measurement of these domains are lacking. To address this deficiency in the research, the International Study of City Youth (ISCY) took on this challenge and developed a framework for measuring secondary school students' non-cognitive skills alongside selected measures of student dispositions and engagement towards school (Lamb, Jackson, & Rumberger, 2015). Moreover, within this project, a larger survey aiming at measuring engagement and non-cognitive skills was developed. In addition, a test for measuring mathematic- and reading competence was distributed together with the survey. In the following, the theoretical underpinnings of selected measures (i.e., academic behavioral skills and engagement) relevant for this validation study are described.

2.1. Academic behavioral skills

Various frameworks emphasized that *21*st *century skills* are comprehensive and include several competence areas and competencies (e.g., Griffin et al., 2012; Voogt & Roblin, 2012). Measuring the full set of 21st century skills in one project may not be possible nor useful, and given the broad set of competence areas it includes, it is both time consuming and challenging (Griffin et al., 2012). Hence, in this paper, we focus on the a*cademic behavioral skills* as the overarching construct, comprising of the three dimensions: *self-efficacy, social skills* and *self-control*. The term academic behavioral skills draws from the research on *Academic mindsets* as an important aspect of deeper learning (Farrington, 2013). Most importantly, previous research has shown that traits such as self-efficacy, communication- and collaboration skills, grit and conscientiousness are positively related to students' academic achievement (Credé et al., 2017; Farrington, 2013). Scholars accentuated that these competencies are malleable (Gutman & Schoon, 2013), highlighting their value in education, and supports the notion of growth mindsets as opposed to fixed mindsets (Dweck, 2006).

2.1.1. Self-efficacy

Self-efficacy is a frequently used measure that is defined as an individual's beliefs about his or her capabilities and levels of performance related to a course of action (Bandura, 1997). Self-efficacy has been widely studied as a model that can explain motivation and behavior, and positively affects academic performance (Caprara et al., 2008; Martin, Montgomery, & Saphian, 2006; Robbins et al., 2004).

2.1.2. Social skills

Social skills include competencies such as communication, collaboration, cooperation, sharing, and helping. Social skills have been defined as "socially acceptable learned behaviors that enable a person to interact effectively with others and to avoid socially unacceptable responses" (Gresham & Elliott, 1990, p. 1). In literature, social skills are addressed as particularly important for future work and life outcomes (Griffin et al., 2012; OECD, 2015). Yet, a direct relation between social skills and academic achievement is considered to be more tenuous (Farrington et al., 2012). However, researchers emphasize that social skills interact with other skills, cross-fertilize and further increase pupils' possibilities to achieve positive outcomes later in life (OECD, 2015, p. 14). Acknowledging that social skills are important for adolescents preparing for adulthood and future workforce, previous research focused mostly on young children (Gutman & Schoon, 2013). Therefore, the focus of this study is on upper secondary students.

2.1.3. Self-control

Self-control is an important aspect of academic perseverance and is understood as the capability to resist short-term impulses to pursuit longer-term goals (Farrington et al., 2012). Researchers have shown that self-control is significantly related to school attainment (Duckworth, Tsukayama, & May, 2010), and higher self-control in childhood is related to better economy and health in adulthood, and lower criminality (Moffitt et al., 2011).

2.2. Engagement

Engagement has also received much attention in education as a solution to the challenges of dropout, low achievement, and alienation

(Fredricks et al., 2004).

Engagement is defined in different ways in the literature (Fredricks et al., 2004) and different concepts are used indistinctly (Libbey, 2004). Boekaerts (2016) described engagement as "a student's active involvement and participation in school-based activities, more concretely it entails students' reactions to and interactions with the learning material as it is embedded in the physical, instructional and social environment" (p. 81). Fredricks et al. (2004) accentuated that there is a consensus in understanding engagement as a multidimensional construct including three dimensions: behavioural-, cognitive-, and emotional engagement. The multidimensionality perspective allows us to analyse the interrelations between the different sub-dimensions of engagement. Even though there is a large agreement on understanding engagement as multidimensional (behavioural, cognitive and emotional), more recent discussions have begun to point to the overlaps between them (Renninger & Hidi, 2016).

Despite the differences in the conceptualization of engagement, there is broad consensus and empirical support regarding the relationship between engagement, academic achievement and school behaviour (Appleton, Christenson, & Furlong, 2008). Several researchers have asserted that school engagement is one of the most important factors associated with school dropout and educational attainment (Jimerson et al., 2003; Te Wang et al., 2019). Also, it is associated with positive academic outcomes (Fredricks et al., 2004).

Behavioral engagement is related to specific student behaviors, such as attendance, behavioural problems and involvement in school or school activities. Young people have a high level of behavioral engagement when they regularly attend class, do not get into trouble, and go to class with a good predisposition to learn (Finn & Voelkl, 1993). These school behaviors are related to successful school performance (Farrington et al., 2012). There are also researchers who measured behavioral engagement or student disengagement based on misbehavior (Janosz et al., 2008), problem behaviors (Palardy, Rumberger, & Butler, 2015) or deviant behaviors (Palardy & Rumberger, 2019).

Cognitive engagement is understood as the level of investment in learning, being thoughtful, strategic and willing to strive for understanding complex ideas and master difficult tasks (Fredricks et al., 2004). In the literature, different ways of measuring cognitive engagement have been utilised. On the one hand, observable aspects such as time-on-task, class participation, completion of homework (Appleton et al., 2008) are often extracted from teachers' observations of students' behaviour in the classroom (Finn & Zimmer, 2012). On the other hand, students' self-reports focusing on schoolwork and persistence when the content is difficult have been used (Finn & Rock, 1997; Reeve & Tseng, 2011). Some researchers also emphasized students' self-regulation of academic effort (Fredricks et al., 2004).

Emotional engagement focuses on interest in school and denotes the extent of positive and negative reactions to teachers, classmates, and school, including a sense of belonging and identification with school and subject domains (Boekaerts, 2016). There are researchers who analysed aspects such as happiness, anxiety, sadness (Skinner, Kindermann, & Furrer, 2009), while others focused on socioemotional aspects of school (Fredricks et al., 2004; Janosz et al., 2008) or school belonging and valuing (Wang, Willett, & Eccles, 2011). In addition, some studies pointed out that high levels of emotional engagement are related to high levels of cognitive engagement and positive behaviour within the school (Rotermund, 2010).

2.3. Correlates of engagement and academic behavioral skills

Because engagement is considered critical to enhancing students' learning outcomes and overcome dropout (Fredricks et al., 2004), it is important to scrutinize the aspects that influence students' school engagement. Several studies have suggested that dispositions such as belonging, self-efficacy, hope, purpose (Lamb et al., 2015) or self-perceptions (Green et al., 2012) are important precursors of students' school engagement and they can also influence cognitive and non-cognitive skills (Lamb et al., 2015). These dispositions are based on students' previous academic and social experiences (Lamb et al., 2015) and can be conditioned by students' sociodemographic characteristics such as socioeconomic status (SES), gender, or being part of an ethnic minority (Bonal, 2003). Moreover, self-concept has been identified as a variable which affects students' performance (Green et al., 2012). A direct effect of self-concept on engagement has been demonstrated; also, an effect of academic self-concept and certain academic dispositions on achievement has been identified (Rotermund, 2010). These investigations conclude that the effects of the academic dispositions or skills on achievement can be mediated by school engagement (Green et al., 2012). However, researchers stressed that although the three types of engagement have been measured largely, there is little documentation of their construct validity, and previous research used engagement as an outcome variable rather than predictor (Gutman & Schoon, 2013).

2.4. Academic performance

Students' academic achievement has often been used in educational research as an outcome variable to identify which variables affect school success (Ross, 2008). Moreover, both self-reported and performance-based measures of students' academic performance are used. Although these approaches are clearly different, researchers accentuate their strengths and weaknesses. Self-reported academic performance is often measured by variables such as self-efficacy, interest, academic aspirations, and/or self-expectations- or teachers' expectations of performance. Moreover, it is considered cheaper and easier to measure along with other more personal variables such as attitudes, perceptions and feelings. Also, it might be prone to bias given students' over- or under-estimation of their own competence. Performance-based measures are more time-consuming and costly in the sense that for instance, teachers and/or others have to develop valid assessments; yet, they oftentimes provide more objective, accurate and reliable measures. In the field of ICT competence (Aesaert, Voogt, Kuiper, & van Braak, 2017), Mathematics (Chen, 2003) and other fields (Bol, Hacker, O'Shea, & Allen, 2005; Miller & Geraci, 2011; Südkamp, Kaiser, & Möller, 2012) researchers have investigated the accuracy and bias of students´ self-reported and performance-based competence. Recognising that such an approach is interesting, it is out of the scope of this study. However, we used both types of measures, i.e., self-reported academic performance (interest in school work, academic aspirations, and teachers' ratings) and performance-based measures (scores on a mathematic- and reading test) to investigate the validity and reliability of the ISCY-survey. Moreover, these measures are further described in the method section.

2.5. The present study

The present study aims at validating the ISCY survey. Following Messick's (1995) conceptualization of validity, we gather evidence on construct validity from two sources. Firstly, we study the factorial structure of the two main constructs, i.e., academic behavioral skills and engagement, as consisting of three dimensions, each as described in the theoretical framework. More specifically, we investigate to what extent the underlying theoretical assumptions of the engagement scale (consisting of the three sub-dimensions: behavioral-, cognitive-, and emotional engagement) and the academic behavioral skills scale (consisting of the sub-dimensions: self-efficacy, social skill and self-control) can be confirmed. Hence, our expectation is that a three-factor measurement model represents the structure of the academic behavioral skills and engagement scales appropriately (internal validity). Secondly, addressing external validity, we investigate the relations between academic behavioral skills and engagement to indicators of students' achievement. Finally, we test for the Differential Item Functioning (DIF) of the survey items across gender, SES, immigrant status

and city background.

Taken together, we address the following three research questions:

1 To what extent can the hypothesized structure of academic behavioral skills (i.e., self-efficacy, social skills and self-control) and engagement (i.e., behavioral-, emotional-, and cognitive engagement) be confirmed? (Internal validity; RQ1).
2 To what extent do the academic behavioral skills and engagement provide invariant measures across students' gender, city, SES, and immigrant background? (Generalizability; RQ2).
3 To what extent are academic behavioral skills and engagement correlated with students' academic performance? (External validity; RQ3).

3. Method

3.1. Sample and procedure

In the present study, we use data from the International Study of City Youth (ISCY[1]) project, which isa longitudinal cross-national study tracking one cohort of students for four or five years in and beyond upper secondary (i.e., high school) in different cities around the world. The sample comprised of $N = 8520$ students (51.6 percent females) in grade 10 (ages 15–16). Data were collected from four large cities in four different countries–Reykjavik in Iceland, Barcelona in Spain, Ghent in Belgium and Bergen in Norway. All schools with students in the target group were invited to participate (except for Barcelona, which applied a two-stage stratified sampling procedure to achieve statistical significance for the whole city of Barcelona given the number of private and public schools and location of the school). Table 1 shows the sample sizes and response rates from each city, including the number of respondents, –and indicates representative samples with response rates ranging between 80 and 92 percents. The ISCY survey and the mathematics and reading tests were administered online, which was supervised by a teacher or a member of the research team. ConQuest 4 (Adams et al., 2012) and Mplus 8 (Muthén & Muthén, 1998–2014Muthén and Muthén, 1998Muthén & Muthén, 1998–2014) were used for statistical and psychometric modeling of the response data.

3.2. Measures

In the present study, we focused on academic behavioral skills, engagement and academic performance scales. These scales consist of items that reflect the intrapersonal, interpersonal and cognitive domains as described by Pellegrino and Hilton (2012) and Palardy and Rumberger (2019). The measures were obtained from self-reports.

3.3. Students' academic behavioral skills

Using 21 items, we assessed the academic behavioral skills construct (Table 2) focusing on: self-efficacy (6 items; e.g., *I am confident of doing well in school*); social skills, which contained items related to communication and collaboration (12 items; e.g., *I work well in groups*); and self-control (3 items; e.g., *I am easily distracted in class*). Students were asked to rate the degree to which they agree on different statements

[1] This article uses data from the International Study of City Youth (see iscy.org). ISCY is an international collaborative project designed and implemented by various research partners from across the world and led by the Centre for International Research on Education Systems (CIRES) at Victoria University, Australia. ISCY has received funding from the Australian Research Council, the Victorian Department of Education and Training and CIRES. The participating countries received funding from their respective national research councils.

Table 1
An overview of the response rates in each city, including the number of invited schools and students and those who participated.

City	% Response rate	N schools participated/ invited	N students participated/ invited
Bergen	80	25/25	2147/2678
Barcelona	92	27/29	2056/2243
Ghent	90	30/39	2354/2608
Reykjavik	81	44/44	1963/2408
Total		126/137	8520/9937

Table 2
An overview of the item wordings and the dimensions which they were allocated to.

Item no	Item wording	Dimension
ST_17_01	I like being at school	Emot.Eng
ST_17_02	I feel safe at school	Emot.Eng
ST_17_07	I get on well with most of my teachers	Emot.Eng
ST_17_11	I will leave this school with good memories	Emot.Eng
ST_17_12	I get a feeling of satisfaction from what I do in class	Emot.Eng
ST_17_13	Working hard in school matters for success in the workforce	Emot.Eng
ST_17_14	What we learn in class is necessary for success in the future	Emot.Eng
ST_17_15	School teaches me valuable skills	Emot.Eng
ST_17_16	My classes give me useful preparation for what I plan to do in life	Emot.Eng
ST_17_06	I find most school work boring	Emot.Eng
ST_17_08	School is often a waste of time	Emot.Eng
ST_22_01	Skipped a class without permission	Behav.Eng
ST_22_02	Been absent from school for a day without permission	Behav.Eng
ST_22_03	Been in trouble with a teacher because of your behaviour	Behav.Eng
ST_22_04	Been given a detention	Behav.Eng
ST_22_05	Arrived late at school	Behav.Eng
ST_17_10	I get into trouble frequently at school	Behav.Eng
ST_51_05	I always get work in on time	Cogn.Eng
ST_26_07	In class, I try to work as hard as possible	Cogn.Eng
ST_26_08	In class, I keep working even if the material is difficult	Cogn.Eng
ST_26_09	In class, I put in my best effort	Cogn.Eng
ST_17_04	I always try to do my best	Cogn.Eng
ST_51_11	I persevere with a job until it is done	Cogn.Eng
ST_51_13	I am a hard working student	Cogn.Eng
ST_32_01	I am confident of doing well in school	Self-efficacy
ST_32_03	Right now I see myself as being pretty successful as a student	Self-efficacy
ST_32_05	I can think of many ways to reach my current goals	Self-efficacy
ST_32_06	There are lots of ways around any problem that I am facing now	Self-efficacy
ST_32_02	I am confident of finding a good job when I finish my studies	Self-efficacy
ST_32_04	There is little that can prevent me from reaching my goals	Self-efficacy
ST_51_01	I work well in groups	Social skills
ST_51_08	I treat others fairly	Social skills
ST_51_03	I understand how others are feeling	Social skills
ST_51_04	I prefer to work alone	Social skills
ST_51_14	I take time to help others	Social skills
ST_51_06	I get along well with others	Social skills
ST_51_18	I am good at getting ideas across in discussions	Social skills
ST_51_09	I express ideas clearly in oral presentations	Social skills
ST_51_10	I am good at leading others	Social skills
ST_51_02	I express ideas clearly in written text	Social skills
ST_51_12	I like to think of new ways to do things	Social skills
ST_51_16	I am good at coming up with new ideas	Social skills
ST_51_07	I tend to leave things to the last minute	Self-control
ST_51_15	I am easily distracted in class	Self-control
ST_51_17	I tend to be lazy	Self-control

Note. Emot.Eng = emotional engagement; Behav.Eng = behavioral engagement; Cogn.Eng = cognitive engagement.

about themselves and their schooling on a 4-point Likert scale (*1 = strongly disagree, 4 = strongly agree*). Note that there were statements that were negatively worded to circumvent and detect random response patterns (e.g., students who selected only one category across all items). These were reverse-coded in the analysis. The item wordings and the dimensions they belong to are provided in Table 2.

3.4. Students' engagement

Using 24 items, we assessed the engagement (see Table 2 for item wordings) construct focusing on the three dimensions: behavioral (6 items; e.g., *Skipped a class without permission*); cognitive (7 items; e.g., *In class, I try to work as hard as possible*) and emotional (11 items; e.g., *I like being at school*). Students were asked to rate the degree to which they agree on different statements about themselves and their schooling on a 4-point Likert scale (1 = *strongly disagree*, 4 = *strongly agree*; 1 = *never*, 4 = *5 times or more*).

3.5. Academic performance

Students' academic performance was assessed using two different approaches. Firstly, we used self-reported measures of: (1) their expectations of academic achievement (i.e., *what results do you expect to get in yourstudies this year?;* 5-point Likert scale);(2) their interest in school work (i.e., *how would you rate your level of interest in school work?*; 5-point Likert scale), and; (3) their perceptions of how their teachers would rate their performance (i.e., *how would your teachers rate you as a student?;* 4-point Likert scale). Secondly, we used performance-based measure of students' scores on the mathematics and reading tests–comprising of 56 and 25 tasks, respectively.

3.6. Student background variables

3.6.1. SES

Students' socio-economic background was obtained by asking students to indicate the highest level of education their mother and father had attained (by selecting between different numbers of categories depending on the national context, and later coded into four categories: no education; lower than ISCED 3 (i.e., upper secondary); ISCED 3; and higher than ISCED 3). Moreover, students were asked to indicate the maternal and paternal employment status by selecting between four categories (1 = working full-time; 2 = working part-time; 3 = not working, but looking for a job; 4 = other, such as home duties, retired). Finally, the number of books at students home was used as a proxy forstudents' social capital (*how many books are there in your home;*coded as 1 = 0–10 books; 2 = 11–25 books; 3 = 26–100 books; 4 = 101–200 books; 5 = 201–500 books; 6 = More than 500 books).

3.6.2. Immigrant status

Students' immigrant background was measured by asking in which country the student, her/his mother and father were born. The responses on the three items were combined to a dichotomous variable indicating immigrant or native status. The student was coded as immigrant if he/she was born abroad, and one of the parents were born abroad, or, if the student was born in the country and one of the parents were born abroad.

3.7. Statistical analyses

Item response theory (IRT) approach (Wilson, 2005) was used to investigate the reliability and validity of the measurement instrument. IRT is considered to be a highly useful methodology for the development of measurement instruments, and specifically for investigating the psychometric quality of such (de Ayala, 2013; Hambleton & Jones, 1993). One of the main advantages of the IRT approach is that it focuses on the distinct items as parts of the full measurement instrument, and not solely on the whole test. This focus on the individual item allows individual items to be evaluated, and consequently revised or removed, if needed, to increase the quality of the measure. Moreover, the underlying trait is emphasized, independent of the actual sample of respondents and items (Thomas, 2011; Wilson, 2005). In addition, IRT allows specification of measurement error for each item and person parameters. This implies that the reliability of a test depends on the unique interaction between the test material and the test taker (de Ayala, 2013). Given the multidimensional nature of the measurands in this study, we compared unidimensional Rasch model with its multidimensional variant (Adams, Wilson, & Wang, 1997). Additionally, exploratory multidimensional IRT approach was used to further scrutinize the a priori specified factorial structure .

3.8. Unidimensional- and multidimensional Rasch models

Unidimensional and multidimensional IRT models are commonly used to measure individuals' latent traits, which cannot be observed directly (Baker & Kim, 2004).

3.8.1. Unidimensional Rasch model

Within IRT, the unidimensional Rasch model (Rasch, 1960) has interpetational advantages over other models and relies on the assumptions of unidimensionality, sufficiency, monotonicity, and local independence (see Wilson & Gochyyev, 2013 for detailed explanations). In Fig. 1, the structure of the unidimensional model–consisting of eight items that load on a single composite factor–is displayed. The estimation of these models is generally done using the marginal maximum likelihood (MML) estimation method. Marginal ML assumes that person-specific parameters are random variables with a particular distribution (Thissen, 1982). To model responses to polytomous items, we used the ordinal variant of the Rasch model the so—called partial credit model (Masters, 1982).

3.8.2. Multidimensional Rasch model

The results obtained from the unidimensional Rasch model are only valid if the assumptions of unidimensionality are satisfied. When a set of items is hypothesized to measure two or more qualitatively distinct latent variables, the multidimensional IRT model should be used instead . The multidimensional random coefficient multinomial logit (MRCML) model was proposed as an overarching Rasch modeling framework focused on analyzing response data from tests and surveys that measure several dimensions (Briggs & Wilson, 2003). By using the MRCML, in addition to item difficulties, the dimension-specific variances and covariances between dimensions can be obtained. Fig. 2 shows the factor structure of the three-dimensional Rasch model, in which we denote correlations between dimensions with z. Note that items can be specified to load on more than a single dimension, as long as conditions for the identification of the model are satisfied.

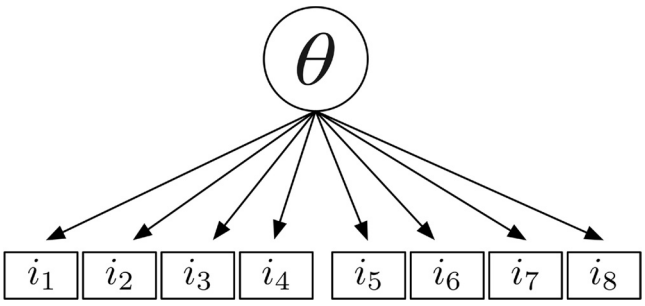

Fig. 1. Unidimensional Rasch model.

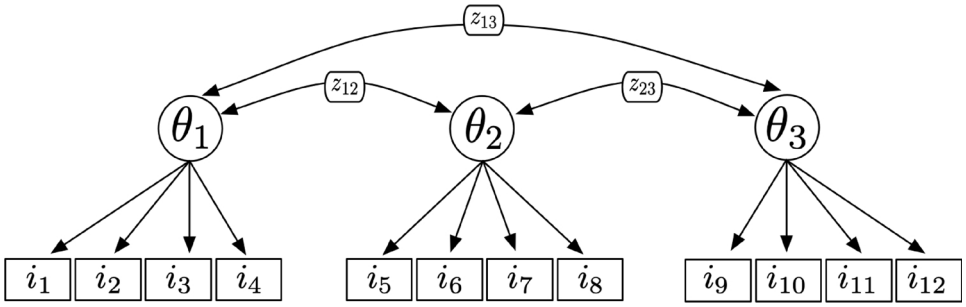

Fig. 2. Three-dimensional Rasch model illustrating the Academic behavioral skills scale (note, for the case of four items per dimension).

3.9. Differential item functioning (DIF)

In this study, we also investigate differential item functioning (DIF) across different demographic groups. The performance gap–on an item or a test– between demographic groups does not imply that an item or a test is biased . Instead, DIF is the existence of performance gaps on a given item after adjusting for potential differences between groups in the overall performance. Specifically, an item exhibits DIF if two students from two different demographic subgroups have the same level of the latent variable but still have different probabilities of answering in the same response category. There are mainly two reasons for conducting DIF-analyses. First, from a measurement development perspective, we aim at ensuring that the measurement instrument itself is not biased towards or against groups (Wilson, 2005) and thus provides evidence for the internal validity of the instrument. Secondly, for drawing valid comparisons across groups, it is critical that the measurement instrument is fair (Millsap, 2011). Thus, testing for DIF offers an additional source of evidence for construct validity (Messick, 1995). Moreover, this is particularly important in large-scale international comparative studies in which the results are used to inform education- and school policies, curriculum development and construction of interventions.

To gather evidence for the internal validity of the instrument across a range of respondent groups (e.g., gender, SES, immigrant- and city background) we investigated whether the items function similarly for different categories of these variables. We view this as the prerequisite for further exploring the external validity (correlations with other variables) and making valid inferences when comparing groups such as males and females. To conduct DIF analysis, we applied an approach commonly used in IRT (Hambleton, Swaminathan, & Rogers, 1991). This approach is based on comparing the relative difficulty of the individual items for each group after accounting for the differences between groups in the overall measure. This approach is easily interpretable and aligns well with the definition that "…an item shows DIF if individuals from different subgroups who have the same ability but they do not have the same probability of getting the item right" (Hambleton et al., 1991, p. 110). For the effect sizes of the DIF statistic, we relied on the recommendation commonly used in the Rasch tradition. Specifically, a statistically significant logit difference value of (1) less than 0.426 is classified as "negligible," (2) between 0.426 and 0.638 is considered "intermediate," and (3) above 0.638 is considered "large" DIF (as suggested by Paek, 2002; described in Wilson, 2005, p. 156).

4. Results

4.1. Internal validity, RQ1

The first research question is concerned with the internal validity of the ISCY survey. We applied Rasch modelling and exploratory factor analysis (EFA) to explore the hypothesized structure of the two scales, namely academic behavioral skills (i.e., self-efficacy, social skills and self-control) and engagement (i.e., behavioral-, emotional-, and cognitive). As shown in Table 3, each of the unidimensional models showed high reliability–0.89 for engagement and 0.82 for academic behavioral skills–estimated with the EAP formulation[2] (Wu, Adams, Wilson, & Haldane, 2007). One item (i.e., *I prefer to work alone*) that was allocated to the social skills dimension was misfitting, and therefore removed, resulting in 20 items in total . We further investigated the proposed multidimensionality of the two scales, each consisting of three dimensions. We compared the unidimensional Rasch model and the three-dimensional Rasch model[3] for each scale by checking whether the difference in deviances between the two models is significant (Rabe-Hesketh & Skrondal, 2005). As shown in Tables 4 and 5, the multidimensional models for academic behavioral skills and engagement scales fit significantly better than their respective unidimensional models. For instance, for the academic behavioral skills scale, the difference in deviances between the two models is 11294.8 (284827 − 273532.2, with 5) degrees of freedom, is statistically significant at 0.01 (based on chi-square distribution). Moreover, standardized root mean square residual (SRMR) shows a good fit for both multidimensional models.– Hu and Bentler (1999) and (Kline, 2005), note that a SRMR value of less than 0.08 is considered a good fit (see Tables 4 and 5).

Thus, we conclude that the three-dimensional Rasch model fits the data statistically significantly better than the simpler unidimensional model. Moreover, the proposed factorial structures with the three factors in each of the two main scales were conceptually sound. These findings were supported by the EFA approach as well.. Specifically, we used EFA on one half of the data to explore whether the hypothesized three-factor structure is recovered. For both of the scales, we found that item allocations are identical to the initially hypothesized structure (for the academic behavioral skills scale, RMSEA = 0.073 and SRMR = 0.042; for the angagement scale, RMSEA = 0.076 and SRMR = 0.044). Next, we used confirmatory approach on the other half of the data and found that model fit indices for both scales are acceptable (for the academic behavioral skills scale, RMSEA = 0.076 and SRMR = 0.055; and for the engagement scale, RMSEA = 0.077 and SRMR = 0.057).

[2] The EAP estimate is a prediction of the respondent's location in the construct, measured based on his/her responses to the relevant set of items. EAP reliability is comparable with the Cronbach's alpha, which is more often used as an index of reliability.

[3] In IRT, the unidimensional Rasch model is a special case of (nested in) the multidimensional Rasch model. The difference in deviances obtained from the estimation of the two models is assumed to have a chi-square distribution, with the difference in the number of parameters as degrees of freedom. Thus, we can statistically test whether the less restricted model (multidimensional Rasch model) fits the data significantly better than the simpler model (unidimensional Rasch model). This likelihood ratio (LR) test can only be used when the models are nested. Note that the dimension-specific variances cannot be nonnegative; thus the null hypothesis is on the boundary of the parameter space, and thus the LR statistic does not have a simple chi-square distribution. The conservative test can then be obtained by simply dividing the naïve p-value from the LR test by 2 (Rabe-Hesketh & Skrondal, 2005).

Table 3
Model fit summaries for Engagement (including 24 items) and Academic behavioral skills (including 21 items).

	Correlation matrix			Reliability	no. of misfitting items	Model fit AIC	BIC
Engagement (unidim)				0.886	0	364397	364911
	Emotional	Behavioral	Cognitive				
Emotional	var: 1.55			0.846	0	352923	353472
Behavioral	corr: 0.54	var: 1.21		0.731	0	352923	353472
Cognitive	corr: 0.71	corr: 0.69	var: 3.14	0.867	0	352923	353472
Academic behavoral skills (unidim)				0.815	1	340695	341146
	Self-efficacy	Social skills	Self-control				
Self-efficacy	var: 3.33			0.826	0	324932	325446
Social skills	corr: 0.61	var: 1.28		0.806	0	324932	325446
Self-control	corr: 0.46	corr: 0.18	var: 1.43	0.654	0	324932	325446

Table 4
Model summaries for the unidimensional and multidimensional Rasch model for Academic behavioral skills.

Academic behavioral skills	Unidimensional model fit	Multidimensional model fit
Deviance	284827	273532.2
Number of persons used	8422	8422
Number of items	18	18
Number of estimated parameters	55	60
Item threshold parameters	54	54
Item slope parameters	0	0
Regression parameters	0	0
AIC	284937	273652
BIC	285324	274075
aBIC	285149	273884
CAIC	285379	274135
Absolute model fit		
Standardized root mean square residual (SRMR)	0,101	0,044

Table 5
Model summaries for the unidimensional and multidimensional Rasch model for Engagement.

Engagement	Unidimensional model fit	Multidimensional model fit
Deviance	336755	327090.1
Number of persons used	8456	8456
Number of items	22	22
Number of estimated parameters	67	72
Item threshold parameters	66	66
Item slope parameters	0	0
Regression parameters	0	0
AIC	336889	327234
BIC	337361	327741
aBIC	337148	327512
CAIC	337428	327813
Absolute model fit		
Standardized root mean square residual (SRMR)	0,075	0,05

Note that RMSEA values less than 0.08 are considered adequate (Bollen, 1989).

We investigated the item fit (i.e., weighted mean square fit statistic) for each item to inspect the alignment of the items with the multidimensional model. Item fit statistic shows the discrepancy between the theoretical and the observed item characteristic curves (Wu & Adams, 2013). Although these values are ideally expected to be close to unity, a common convention is that values between 3/4 (0.75) and 4/3 (1.33) are acceptable lower and upper bounds (Adams & Khoo, 1996). For both scales–academic behavioral skills and engagement–none of the items were misfitting.

4.2. Generalizability, RQ2

To investigate the extent to which the scales function equally across demographic groups, we conducted DIF-analyses with respect to students' gender, SES, immigration status and their city. Our findings showed that two items from each scale were flagged for DIF. In particular, the item "*Right now I see myself as being pretty successful as a student*" was found to be biased against students with parents with no education and students from Barcelona, while it favored Reykjavik students. Moreover, the item "*There is little that can prevent me from reaching my goals*" favored students with low SES (mother with no education). These two biased items were removed, leaving the academic behavioral scale with 18 items. From the engagement scale, the item "*My classes give me useful preparation for what I plan to do in life*" (emotional engagement) favored Barcelona students. In addition, the item "b*een given a detention*" (behavioural engagement) favored students from Bergen and Reykjavik, and was biased against students from Barcelona. These two items were removed, leaving 22 items representing the engagement scale. It is challenging to explain why these items exhibited DIF. One explanation regarding the items favoring some cities is that the translation might have slightly changed the interpretation of what is being asked in those items. Hence, we suggest that native language experts inspect these items to identify potential factors for DIF. Moreover, in the future administrations of this survey, we recommend revising these items.

After removing the four DIF items, we again conducted DIF analyses, which showed that none of the items were flagged for DIF. In summary, four out of 44 items exhibited bias and were removed. We believe this strengthens the generalizability of the survey, and allows valid comparisons with respect to gender, SES, immigration status and city.

Using the remaining 40 items (18 representing the academic behavioral skills, and 22 representing the engagement scales), we confirmed our findings with regards to the dimensionality . Further, the item fit analysis showed that none of the 40 items fell outside of the acceptable range (Appendix A, Table A1).

The estimated correlations among the three dimensions of the academic behavioral skills (i.e., *self-efficacy, social skills and self-control*) varied between 0.18 and 0.61 (see Table 6). Social skills and self-

Table 6
The reliability of and correlations between the three dimensions of academic behavioral skills.

	Self-efficacy	Social skills	Self-control
Self-efficacy			
Social skills	0.611		
Self-control	0.464	0.179	
Variances	3.356	1.282	1.345
EAP Reliability	0.771	0.806	0.639
Cronbach's alpha	0.783	0.823	0.659

Table 7
The reliability of and correlations between the three dimensions of engagement.

	Emot.Eng	Behav.Eng	Cogn.Eng
Emot.Eng			
Behav.Eng	0.565		
Cogn.Eng	0.719	0.688	
Variances	1.47	1.206	3.138
EAP Reliability	0.833	0.714	0.866
Cronbach's alpha	0.83	0.711	0.869

Note. Emot.Eng = emotional engagement; Behav. Eng = behavioral engagement; Cogn.Eng = cognitive engagement.

control showed the lowest correlation (0.18), while the highest correlation was between self-efficacy and social-skills (0.61). The correlations among the three dimensions of the engagement scale (*behavioral, emotional and cognitive engagement*) varied between 0.57 and 0.72 as shown in Table 7.

The reliabilities (both EAP reliability and Cronbach's alpha) for each dimension in the academic behavioral skills and engagement scales are shown in Tables 6 and 7, respectively. According to the rules described by DeVellis (2012): Cronbach's alpha between .65 and .70 is minimally acceptable; Cronbach's alpha between .70 and .80 is respectable; and Cronbach's alpha > .80 is very good. Most of the dimensions show acceptable levels of reliability, with the lowest reliability estimated for the self-control dimension (Cronbach's alpha = 0.66; Table 6). This is likely due to the low number of items (i.e., three items) measuring this dimension, and, if possible, we recommend adding one or two items to this sub-scale in the next cycle of the survey.

The item difficulties, and item and category fit statistics (Appendix A, Tables A1 and A2) indicate that all items and steps fit well . Moreover, no redundant items were identified, as shown in Appendix A, Tables A3 and A4.

4.3. External validity, RQ3

We tested the predictive power of the three engagement and academic behavioral skills dimensions in predicting academic–reading and mathematics–outcomes. Reading scores ranged from 1 to 24, with the mean score of 11.70 and the standard deviation of 4.14. Mathematics scores ranged from 1 to 30, with the mean score of 12.75 and the standard deviation of 4.56.

4.4. Engagement

Among the three engagement dimensions, we found that the behavioral engagement is the strongest predictor of the reading and mathematics scores. In particular, controlling for the other two engagement dimensions (i.e., cognitive and emotional), we found that for every level increase in behavioral engagement (i.e., going from "1 or 2 times" to "Never") on a single item (out of five behavioral engagement items), the reading score is estimated to increase by about 0.22 points—and this increase is significant at 0.0001 level. This implies that if a student responds one level higher in each of the five items, that students' reading score is expected to increase by about a quarter of standard deviation (i.e., one point in the reading sum score).

The similar pattern is observed when we try to predict the mathematics score . Controlling for the other two engagement dimensions, we found that for every level increase in behavioral engagement, the mathematics score is estimated to increase by about 0.17 points, which is statistically significant at 0.0001 level. To put this in perspective, this implies that when comparing two students, the one who responded one level higher on each of the five behavioral engagement items is estimated to have about 20 % standard deviations higher in mathematics score (i.e., about 0.85 points higher).

Moreover, we attempted to predict students' self-reported ratings of: (1) how their teachers would evaluate their competence (range = 1–4; mean = 2.55; SD = 0.80); (2) their interest in school work (range 1–5; mean = 3.07; SD = 0.93), and; (3) what results they expect to get this year in their studies (range 1–5; mean = 3.73; SD = 0.80).

When predicting student's expectation of their teacher's rating, we found that cognitive engagement is by far the best predictor (compared to other two engagement dimensions)—statistically significant at 0.0001 level . Although the other two dimensions also did predict this statistically significantly, the effect sizes were low . We found that for every category/level increase in any of the cognitive engagement items (out of seven items), student's expectation of his/her teacher's rating increases by about 0.07 points—or by about 9% standard deviations. We also found that these three engagement dimensions explain about 20 % of the variation on that question.

Cognitive engagement was also the only predictor of students' rating of their interest in schoolwork (controlling for the other two engagement dimensions). We found that for every level increase in any of the cognitive engagement items, student's rating of their interest in schoolwork is expected to increase by about 0.016 points (or by about .02 standard deviations), significant at 0.0001 level, indicating a small effect in terms of practical significance.

Cognitive engagement was also the best predictor of students' expectations from studies for the year. For one level increase in any of the seven cognitive engagement items, the student's expectation for the year is estimated to increase by about 9 % of the standard deviation (statistically significant at 0.0001 level). Moreover, we found that these three engagement dimensions explain about a quarter (R-squared = 0.26) of the variance in the students' rating of their expected performance for the school year.

4.5. Academic behavioral skills

Among the three disposition dimensions, self-efficacy is the best predictor of reading and mathematics scores. For every category increase in any of the four self-efficacy items, we found that the reading score is expected to increase by about 0.18 points (or about 4 % standard deviation), and the mathematics score is expected to increase by about 0.19 (about 4 % standard deviation)—both of which are statistically significant at 0.0001 level.

In addition, we found that all the three academic behavioral skills dimensions are statistically significant predictors of students' expectation of their teachers' rating. In terms of practical significance, we found that self-efficacy has the highest effect size among the three. Specifically, for every level increase in any of the self-efficacy items, students' expectation of their teachers' rating is estimated to increase by about 0.14 standard deviations. Self-control was also found to be a good predictor of student's perception of his/her teacher's rating. For every unit increase in any of the self-control items, typical student's rating is expected to increase by about 0.10 standard deviations. We also found that the three academic behavioral skills dimensions explain about 21 % of the variation in the outcome variable (students' expectation of their teachers' ratings).

We did not find supporting evidence that academic behavioral skills dimensions predict students' rating of their interest—even though self-efficacy was statistically significant at 0.004 level but practically insignificant (i.e., only about 2 % of standard deviation change in the outcome variable for every unit increase in self-efficacy).

Finally, we found that self-efficacy and self-control are good predictors of students' expectations for the year, with 17 % and 10 % increase in the outcome variable for every unit increase in self-efficacy and self-control, respectively. We also found that the three academic behavioral skills dimensions explain about 25 % of the variation in students' expectations for the year.

5. Discussion

5.1. Internal validity, RQ1

Our findings regarding the internal validity of the ISCY survey showed that one of the 45 items were misfitting, and therefore removed for the purposes of this paper. The misfit of this specific item implied that the item might not be aligned well with the other items with regards to dimensionality. Still, we support that a revised version of the item should be administered in a future cycle of the survey . Evidence regarding sufficient levels of reliability was obtained for both scales (EAP reliability estimated at 0.82 for the academic behavioral skills scale and at 0.89 for the engagement scale). The reliability of the engagement scale in this study lies in the upper half when compared to previous research, which identified reliabilities of several measures in the range of .54 to .93, and most scales lying between .70 and .80 (Fredricks & McColskey, 2012).

Further, we investigated whether the proposed factorial structure of the two main scales is justified. Our findings showed that the multidimensional model–when applied on academic behavioral skills and engagement scales (each consisting of three dimensions), –fits better than the unidimensional models. These findings lend support to the internal validity of the measures and add evidence to the robustness of the factorial structure for the academic behavioral skills (Griffin et al., 2012) and engagement (Fredricks et al., 2004) scales .

Our findings on the structure of academic behavioral skills are in line with previous research, which used performance-based test and identified similar aspects including social skills (i.e., communication and collaboration) and self-efficacy (Siddiq, Gochyyev, & Wilson, 2017).

Engagement has been studied in different areas and taking several perspectives, but there is little consensus regarding its construction. For this reason, it is important to deepen its construction and validation. The results of this study are consistent with recent literature on school engagement, supporting the theory that school engagement is a multidimensional construct consisting of the three dimensions: behavioural, cognitive and emotional engagement (Fredricks et al., 2004). The multidimensional structure of the engagement construct allows us to analyse different elements of engagement and assess which specific dimension of engagement has the greatest impact on different outcomes. Such knowledge is of interest as it can contribute to developing targeted school interventions, which focuses on the specific types of engagement and not only on engagement in general.

5.2. Generalizability, RQ2

The results obtained regarding the DIF analysis revealed that none of the items are biased towards students' gender and immigrant background. Of the four DIF items, only two showed DIF across students' SES, and three were biased with respect to cities. Consequently, the four items that exhibited DIF were removed. This finding lends support to the internal validity of the survey and ensures that the measure is fair across students' gender, SES, immigrant- and city background – further suggesting that valid comparisons between these groups can be made.

DIF analyses are increasingly important – particularly in cross-national studies (Dorans & Holland, 1993). As reported in PISA: "The interpretation of a scale can be severely biased by unstable item characteristics from one country to the next" (OECD, 2006, p. 86). In these cases, the students' responses to items might suffer from bias due to translation into several languages or reflect differences in national curricula rather than students' content domain ability. Moreover, it is vital to carry out analyses to validate such questionnaires because several researchers have pointed out that there are, for instance, important differences between engagement of students depending on whether they belong to an ethnic minority or not (Fredricks et al., 2004). However, only few previous studies have investigated the validity of the instruments regarding their fairness across different groups. Therefore, it is also important to check whether the different facets that measure academic behavioural skills and engagement operate in a similar way for the different socio-demographic profiles in the sample. Once this check has been carried out, providing evidence for the fairness of the measurement instrument, the effects of these constructs on, for example, the school performance of certain groups can be analysed.

5.3. External validity, RQ3

The results of the relations between student achievement (both self-reported and performance-based) and academic behavioral skills and engagement showed positive and significant correlations. These findings augment the validity argument for the ISCY survey as the positive correlations support the external validity. This is in accordance with previous literaturewhich showed that increased engagement and academic behavioral skills are related to positive school performance (Pellegrino & Hilto, 2012; Gutman & Schoon, 2013).

Our findings reveal that for each unit increase in the behavioral engagement level (controlling for the two other engagement dimensions) the reading- and math scores increase significantly—behavioral engagement being the best predictor. In this regard, the results of previous research are indistinct. The effects of behavioral engagement on academic performance varies depending on how performance is measured (e.g., teacher grades, external performance tests) and how engagement is defined and constructed (Fredricks et al., 2004). Hence, further research is needed to explore the three engagement dimensions and their predictive value.

In this study, we supplemented the performance-based outcomes with more subjective performance indicators (e.g., the students' perceptions regarding how their teachers rate them, what interest they have in schoolwork and their performance expectations). From these analysis, we observed that cognitive engagement plays the most important role, This implies that students who see themselves as hard-working think that they will obtain good grades and a good evaluation by their teachers.

Further, our results do not show a significant effect of emotional engagement on the analyzed outcome variables. In fact, this is in line with previous research, showing that emotional engagement is not a strong predictor of academic performance (Finn & Voelkl, 1993). However, there are studies that showed a positive relation between emotional engagement and academic performance with the mediation of behavioral engagement (Lee, 2014), and that high levels of emotional engagement are related to higher levels of behavioral and cognitive engagement (Rotermund, 2010).

Finally, it is important to mention that the engagement of students in school is influenced by individual and family factors (e.g., gender, grade level, ethnicity, language spoken at home, socioeconomic level) and school characteristics (e.g., percentage of ethnic minority students, public or private). For this reason, in future studies, it would be vital to examine whether the effects of the different dimensions of engagement and academic behavioral skills vary depending on schools and students' profiles.

In sum, this study supplements the literature on the measurement and validity of academic behavioral skills and engagement – an aspect that has rarely been studied in previous literature (Glanville & Wildhagen, 2007; Appleton, 2008; Gutman and Schoon, 2013). Another important contribution of this article is the analysis of the invariance across gender, SES, immigrant and city- background, because little research compared the effects of academic behavioral skills and engagement across different student backgrounds.

5.4. Limitations and future directions

The data used in this study are mainly based on students' self-

reports, and therefore, their responses may be conditioned by external social aspects depending on what they believe is expected from them. This could introduce bias into the results (Wang et al., 2011). Future research should consider additional sources such as teacher-reports or school record data. Second, the analysis carried out do not have a longitudinal design. For this reason, future studies may benefit from including perspectives related to changes in academic behavioral skills and engagement across time, and the relations between the two and academic performance. Specifically, intervention studies should focus not only on measuring students' academic behavioral skills and engagement, but also on how they can be trained and implemented in schools and curricula.

5.5. Conclusion and implications

In the last few decades, there has been an increased emphasis on the prominence of certain academic skills and engagement in improving students' academic performance and reducing dropout rates. It has been argued that the dispositions or behaviors that have been analyzed in this article are malleable (Gutman & Schoon, 2013), signifying that they can be developed and transformed. Aligning with this view, our results show that the teachers and the schools could benefit from supporting the students' academic behavioral skills and their engagement to increase academic performance.

We believe these results will help to deepen the debate on the construction and validation of academic behavioral skills and engagement scales, and provide knowledge about their role in explaining academic performance. Nevertheless, we acknowledge that further research is needed to explore other aspects related to academic behavioral skills and engagement, for example what factors have greater influence in their development or what role the sociodemographic characteristics, teachers and schools play in the development of such characteristics. We believe this study has shown promising approaches, and laid ground for similar investigations in future studies.

Appendix B. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.stueduc.2020.100880.

References

Adams, R. J., & Khoo, S.-T. (1996). *Quest.* Melbourne, Australia: ACER Press [computer program].

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

AERA, APA, & NCME (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Aesaert, K., Voogt, J., Kuiper, E., & van Braak, J. (2017). Accuracy and bias of ICT self-efficacy: An empirical study into students' over- and underestimation of their ICT competences. *Computers in Human Behavior, 75*, 92–102. https://doi.org/10.1016/j.chb.2017.05.010.

Appleton, J., Christenson, S., & Furlong, M. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools, 45*, 369–386. https://doi.org/10.1002/pits.20303.

Baker, F. B., & Kim, S. H. (2004). *Item response theory. Parameter estimation techniques* (2nd ed.). Marcel Dekker.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* Freeman.

Boekaerts, M. (2016). Engagement as an inherent aspect of the learning process. *Learning and Instruction, 43*, 76–83. https://doi.org/10.1016/j.learninstruc.2016.02.001.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and exploratory style on calibration accuracy and performance. *Journal of Experimental Education, 73*(4), 269–290.

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

Bonal, X. (2003). *Apropiaciones escolares: entre las posiciones sociales y las disposiciones educativas.* Fundació Jaume Bofill.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87–100.

Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Barbaranelli, C., et al. (2008). Longitudinal Analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology, 100*, 525–534.

Chen, P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*, 79–92.

Clark, K. N., & Malecki, C. K. (2019). Academic Grit Scale: Psychometric properties and associations with achievement and life satisfaction. *Journal of School Psychology, 72*, 49–66. https://doi.org/10.1016/j.jsp.2018.12.001.

Credé, M., Tynan, M., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology, 113*(3), 492–511. https://doi.org/10.1037/pspp0000102.

de Ayala, R. J. (2013). The IRT tradition and its applications. In T. D. Little (Vol. Ed.), *The Oxford handbook of quantitative methods: Vol. 1*, (pp. 144–169). Oxford University Press.

DeVellis, R. F. (2012). *Scale development: Theory and applications.* Sage Publications.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.). *Differential item functioning* (pp. 35–66). Lawrence Erlbaum Associates.

Duckworth, A. L., Tsukayama, E., & May, H. (2010). Establishing causality using longitudinal hierarchical linear modeling: An illustration predicting achievement from self-control. *Social Psychological and Personality Science, 1*(4), 311–317.

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*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087.

Dweck, C. S. (2006). *Mindset: The new psychology of success.* Random House Publishing.

Farrington, C. A. (2013). *Academic mindsets as a critical component of deeper learning.* University of Chicago: Consortium on Chicago School Research. A White Paper prepared for the William and Flora Hewlett Foundationhttps://www.hewlett.org/wp-content/uploads/2016/08/Academic_Mindsets_as_a_Critical_Component_of_Deeper_Learning_CAMILLE_FARRINGTON_April_20_2013.pdf.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., et al. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review.* Chicago: University of Chicago Consortium on Chicago School Research.

Finn, J., & Rock, D. (1997). Academic success among students at risk for school failure. *The Journal of Applied Psychology, 82*(2), 221–234. https://doi.org/10.1037/0021-9010.82.2.221.

Finn, J., & Voelkl, K. (1993). School characteristics related to student engagement. *The Journal of Negro Education.* https://doi.org/10.2307/2295464.

Finn, J., & Zimmer, K. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.). *Handbook of research on student engagement* (pp. 97–131). Springer Science + Business Media. https://doi.org/10.1007/978-1-4614-2018-7_5.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.). *Handbook of research on student engagement* (pp. 763–782). Springer Science + Business Media. https://doi.org/10.1007/978-1-4614-2018-7_37.

Fredricks, J., Blumenfeld, P., & Paris, A. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059.

Glanville, L., & Wildhagen, T. (2007). The measurement of school engagement. Assessing dimensionality and measurement invariance across race and ethnicity. *Educational and Psychological Measurement, 67*(6), 1019–1041. https://doi.org/10.1177/0013164406299126.

Green, J., Liem, G., Martin, A., Colmar, S., Marsh, H., & McInerney, D. (2012). Academic motivation, self-concept, engagement, and performance in high school: Key processes from a longitudinal perspective. *Journal of Adolescence, 35*, 1111–1122. https://doi.org/10.1016/j.adolescence.2012.02.016.

Gresham, F. M., & Elliott, S. N. (1990). *The social skills rating system.* American Guidance Service.

Griffin, P., & Care, E. (2015). The ATC21S method. In P. Griffin, & E. Care (Eds.). *Assessment and teaching of 21st century skills: Methods and approaches* (pp. 3–33). Springer.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills.* Springer.

Gutman, L. M., & Schoon, I. (2013). *The impact of non-cognitive skills on outcomes for young people, Literature review.* The Education Endowment Foundation (EEF). Institute of educationhttps://educationendowmentfoundation.org.uk/public/files/Publications/EEF_Lit_Review_Non-CognitiveSkills.pdf.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement Issues and Practice, 12*, 535–556.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Janosz, M., Archambault, I., Morizot, J., & Pagani, L. (2008). School engagement trajectories and their differential predictive relations to dropout. *The Journal of Social Issues, 64*(1), 21–40. https://doi.org/10.1111/j.1540-4560.2008.00546.x.

Jimerson, S., Campos, E., & Green, J. (2003). Toward an understanding of definitions and measures of school engagement and related terms. *The California School Psychologist: CASP / California Association of School Psychologists, 8*, 7–27. https://doi.org/10.1007/BF03340893.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.). *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Khine, M. S., & Areepattamannil, S. (2016). *Non-cognitive skills and factors in educational attainment. Contemporary approaches to research in learning innovations.* Springer.

Kline, R. B. (2005). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Lamb, S., Jackson, J., & Rumberger, R. (2015). *ISCY technical paper: Measuring 21st century skills in ISCY*Technical Report. Melbourne, Victoria: Victoria University, Centre for International Research on Educational Systems.

Lee, J. S. (2014). The relationship between student engagement and academic performance: Is it a myth or reality? *The Journal of Educational Research, 107*(3), 177–185. https://doi.org/10.1080/00220671.2013.807491.

Libbey, H. (2004). Measuring student relationships to school: Attachment, bonding, connectedness, and engagement. *The Journal of School Health, 74*(7), 274–283. https://doi.org/10.1111/j.1746-1561.2004.tb08284.x.

Martin, J. H., Montgomery, R. L., & Saphian, D. (2006). Personality, achievement test scores, and high school percentile as predictors of academic performance across four years of coursework. *Journal of Research in Personality, 40*, 424–431.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology Learning, Memory, and Cognition, 37*(2), 502–506. https://doi.org/10.1037/a0021802.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* Routledge.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences, 108*(7), 2693–2698.

Muthén, B., & Muthén, L. (2014). *Mplus 7.3.* [Computer Software]Muthén & Muthén.

OECD (2015). *Skills for social progress: The power of social and emotional skills. OECD skills studies.* OECD Publishinghttps://www.oecd-ilibrary.org/education/skills-for-social-progress_9789264226159-en.

OECD (2006). *PISA 2006*Technical report. OECDhttp://www.oecd.org/dataoecd/0/47/42025182.pdf.

Palardy, G. J., & Rumberger, R. W. (2019). The effects of 21st century skills on behavioral disengagement in Sacramento high schools. In J. Demanet, & M. Van Houtte (Eds.). *Resisting education: A cross-national study on systems and school effects* (pp. 53–80). Springer.

Palardy, G. J., Rumberger, R. W., & Butler, T. (2015). The effect of high school socio-economic, racial, and linguistic segregation on academic performance and school behaviors. *Teachers College Record, 117*(12), 1–52.

Partnership for 21st Century Skills (P21) and AACTE (American Association of Colleges of Teacher Education) (2010). *21st century knowledge and skills in educator preparation.* White paperhttp://www.p21.org/storage/documents/aacte_p21_whitepaper2010.pdf.

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Research Council. The National Academies Press.

Rabe-Hesketh, S., & Skrondal, A. (2005). *Multilevel and longitudinal modeling using STATA.* Stata Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Danish Institute for Educational Research. MESA Press (Expanded edition 1983. ed.).

Reeve, J., & Tseng, C. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemporary Educational Psychology, 36*(4), 257–267. https://doi.org/10.1016/j.cedpsych.2011.05.002.

Renninger, K. A., & Hidi, S. E. (2016). *The power of interest for motivation and engagement.* Routledge/Taylor & Francis Group.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261–288.

Ross, S. (2008). *Motivation correlates of academic achievement: Exploring how motivation influences academic achievement in the PISA 2003 dataset.* [Doctoral dissertation, University of Victoria]http://dspace.library.uvic.ca/bitstream/handle/1828/3209/Ross_Shelley_PhD_2008.pdf?sequence=1&isAllowed=y.

Rotermund, R. (2010). *The role of psychological antecedents and student engagement in a process model of high school dropout.* [Doctoral dissertation, University of California Santa Barbara]https://eric.ed.gov/?id=ED522153.

Siddiq, F., Gochyyev, P., & Wilson, M. (2017). Learning in digital networks – ICT literacy: A novel assessment of students' 21st century skills. *Computers & Education, 109*, 11–37. https://doi.org/10.1016/j.compedu.2017.01.014.

Skinner, E., Kindermann, T., & Furrer, C. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement, 69*(3), 493–525. https://doi.org/10.1177/0013164408323233.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627.

Te Wang, M., Fredricks, J., Ye, F., Hofkens, T., & Linn, J. S. (2019). Conceptualization and assessment of adolescents' engagement and disengagement in school: A multidimensional school engagement scale. *European Journal of Psychological Assessment, 35*(4), 592–606. https://doi.org/10.1027/1015-5759/a000431.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment, 18*, 291–307.

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies, 44*, 299–332.

Wang, M., Willett, J., & Eccles, J. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology, 49*(4), 465–480. https://doi.org/10.1016/j.jsp.2011.04.001.

Wilson, M. (2005). *Constructing Measures: An item response modeling approach.* Lawrence Erlbaum Associates, Inc.

Wilson, M., & Gochyyev, P. (2013). Psychometrics. In T. Teo (Ed.). *Handbook of quantitative methods for educational research* (pp. 1–53). Springer.

Wu, M. L., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*(4), 339–355.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *ConQuest: Generalised item response modelling software (Version 2.0).* ACER Press.