Development and Validation of the UiL-Scales for Measurement of Development in Life Skills-A Test Battery of Non-Cognitive Skills for Danish School Children

Author

Makransky, Guido, Wandall, Jakob, Madsen, Simon Ryberg, Hood, Michelle, Creed, Peter

Published

2019

Journal Title

SCANDINAVIAN JOURNAL OF EDUCATIONAL RESEARCH

Version

Accepted Manuscript (AM)

DOI

10.1080/00313831.2019.1595716

Development and Validation of the UiL-Scales for measurement of Development in Life Skills - A Test Battery of Non-cognitive skills for Danish School Children

Guido Makransky[1], Jakob Wandall[2], Simon Ryberg Madsen[3], Michelle Hood[4], and Peter Creed[4].

1. University of Copenhagen
2. NordicMetrics & Aarhus University
3. Epinion
4. Griffith University

Corresponding Author: Guido Makransky, University of Copenhagen

Address: Øster Farimagsgade 2A, DK-1353 Copenhagen K, Denmark (45)31338511

Abstract

Abundant research has pointed to the importance of non-cognitive skills for success in life. This paper describes the development and validation of the "UiL", designed to measure 19 non-cognitive skills that have been identified as being important for school children in Denmark. First, we describe the development of the scales, and then report a preliminary validation with 1560 students from Grades 4 to 9. Second, we present a validation of the final UiL, which was revised and re-administered to the same sample of students. The final validation sample consisted of 1373 students (48.6% boys, ages 9 to 16 years). Results from the Confirmatory Factor Analysis indicated that the UiL had acceptable discriminant and convergent validity. The results from a RASCH partial credit analysis indicated that four of the 19 had excellent fit, with other scales needing nuanced interpretation because of some item misfit, local dependence, multidimensionality, and DIF by grade.

Key words: non-cognitive skills; validation; partial credit model; confirmatory factor analysis

Introduction

In most countries, the primary purpose of basic school education is to develop cognitive knowledge and skills in subjects like reading, mathematics, and science (Bonell, 2014; Heckmann, 2008; Levin, 2012; Pellegrino, 2012). The statement of purpose of the Danish Act of Primary and Lower Secondary Schools (see Appendix 1), emphasizes that cognitive knowledge and skills (e.g., reading and mathematics) are a means to the development of more fundamental characteristics such as motivation for learning, self-efficacy, awareness, imagination, self-confidence, and being a responsible, democratic citizen. Many of these factors are referred to as non-cognitive skills, which have been found to be important predictors of long-term educational outcomes (Heckman & Kautz, 2012). Several studies have shown that educational interventions that target non-cognitive skills in early childhood (e.g., the Perry Preschool Project; Heckman, 2008) can have a positive effect on a wide range of adult outcomes (e.g., crime, education, employment, and health; Heckman et al., 2010). Research also shows associations between perseverance and achieving higher levels of education (e.g., Duckworth, 2007), and conscientiousness is related to better job performance (e.g., Barrick & Mount, 1993), self-control is related to lower criminality and higher academic performance (e.g., Duckworth & Seligman, 2005; Hirschi & Gottfredson, 1995), and well-being is related to better general happiness (e.g., Diener et al., 1999). Further, research has pointed to the malleability of these skills, which means that it is possible to develop these skills through educational interventions (Kautz et al., 2014).

The measurement of non-cognitive skills was an important component in a wide range of evaluation activities during the 20th century (e.g., job selection, skills training, and psychological interventions) and an object of research in the social and educational sciences (Kautz et al., 2014; Heckman & Kautz, 2012). However, solid evidence about the individual student's level of these skills has not been widely accessible for teachers, even though these skills are considered to be as

important as cognitive skills (e.g., intelligence, knowledge and skills in reading and mathematics; Heckmann, 2008). Danish schools/teachers have recently obtained scientifically validated tools to measure cognitive knowledge and skills (e.g., the Danish National Adaptive Test - DNT). The DNT was developed in response to mediocre international test results and recommendations by the OECD that Denmark improve its capacity to evaluate educational outcomes (Mortimore, 2004). The development of the DNT has raised political and public concern that the test would result in a stronger focus on a more narrow aspect of the objectives of the Danish public school system (i.e., Danish and English languages, mathematics, and science). In the legislative process, the government noted that measures of other constructs should be developed. In response to this call, the development of the UiL (*Udviklings i Livsfærdigheder,* translated as *Development in Life Skills*) scales was initiated in 2013 due to a lack of existing validated instruments for measuring non-cognitive skills. The research team collaborated with the two largest Danish municipalities to develop a tool that would provide data to help teachers work with students' non-cognitive skills in a systematic way.

A combination of community needs, backed by recent international research and a renewed political focus on non-cognitive skills, has played a key role in initiating the development the UiL test battery. Although there are several studies that describe the validation of scales to measure a single non-cognitive skill (e.g., Duckworth & Quinn, 2009), and studies that use IRT to validate non-cognitive skill scales (e.g., Ambiel et al., 2015; Stump, 2010), we could find no studies that systematically assessed the validity of a non-cognitive test battery such as the UiL. This article describes the development and validation of the UiL battery, which was designed to measure a wide range of non-cognitive skills that have been identified as being central to primary and lower secondary schools in Denmark.

We provide a brief overview of the development of the UiL, and describe a study that assesses its validity and reliability. Two main research questions were investigated: (a) Do the 19 scales in the UiL have convergent and discriminant validity? and (b) Do the 19 scales have acceptable validity, internal reliability, and measurement invariance across grade levels? The first research question was investigated by testing the fit of data to a 19 factor model using confirmatory factor analysis (CFA; Brown, 2015). The second research question was investigated by testing the fit of each of the 19 scales to the partial credit model (PCM; Masters, 1982) within the framework of item response theory (IRT).

Development of the UiL

The preliminary version of the UiL battery was developed by following the standards for educational and psychological testing from the American Psychological Association standards (Eignor, 2013). To achieve this, we used both qualitative and quantitative methods to identify, assess, and validate the items and scales that made up the final battery.

Method

Participants and Qualitative Procedures

A panel of five international experts from different relevant areas (including psychology, psychometrics, teaching, and economics) worked together with Danish researchers, consultants, and employees (teachers, pedagogical consultants from municipalities) from the two largest municipalities in Denmark (Copenhagen and Aarhus) to develop a list of non-cognitive skills that would be measured in the UiL. The statement of purpose from the Act of Primary and Lower Secondary Schools was used as a reference point (see Appendix 1). The panel of experts then met with stakeholders to analyze which skills aligned with Danish public school values. Literature searches were conducted to identify existing constructs that matched these identified skills. For example, the statement of purpose indicates that “The school must develop working methods and

provide a framework for experience, reflection, and dynamism so students develop awareness and imagination and confidence in their own ability, take a stand, and take action"; thus, non-cognitive skills such as creativity, critical thinking, self-efficacy, self-esteem, and drive were considered relevant to measure.

Several panel discussions were conducted with researchers and stakeholders from the two municipalities to ensure that the non-cognitive skills were relevant for Danish students and teachers, and had solid international evidence of significance for long-term educational outcomes. As a result, the panel identified 17 non-cognitive skills (see Table 1). A literature review was then conducted to find the most relevant existing literature on each of the non-cognitive skills. From these scales, we selected items that were most relevant and suitable to be adapted for the Danish public school context.

The International Test Commission guidelines on test adaptation were used to make the scales suitable for use in Denmark (Hambleton, Merenda, & Spielberger, 2004). Items were translated into Danish and otherwise adapted when they needed to be made relevant for Danish students. Some adaptation was required as some concepts had different connotations in Danish. For example: "Hard working" is an attitude that most American parents want their children to develop, admire, and respect. However, to most Danes, working hard is not admirable – and especially not for children, where independence and tolerance are more desired qualities (Inglehart, 2000). Therefore, more tailored concept descriptions and items were required (e.g., "I continue to work at a problem, even when I fail in my first attempt"). The items were then (a) reviewed by the panel of experts, (b) proofread by pedagogical consultants from the municipalities for content, item relevance, language, and spelling, (c) evaluated in a small pilot with twenty-five 5th grade students in one of the target schools, (d) revised by the research group following student feedback regarding word and sentence difficulty and length of survey, and (e) included in a questionnaire

with a 5-point Likert scale that ranged from *completely agree* (1) to *completely disagree* (5), to be administered to a large sample of Danish school children .

---Insert Table 1 here---

Participants and Quantitative Procedures

The sample consisted of complete results from 1560 students (50.4% boys) from 4th to 9th grades (ages 9 to 16 years) from 8 public schools (4 each from municipalities of Aarhus and Copenhagen). Data were collected in October and November, 2016. All class-teachers and students volunteered to complete the questionnaire. Students were asked to respond to all scales (108 items), in addition to evaluative questions, administered in two sittings due to the large number of items (student responses were matched using student ID). Data from approximately 50 respondents were discarded because students answered fewer than 10 items. Responses were included in analyses when all items for a scale had been completed, but it was not a requirement to have responded to all scales.

Each of the 17 scales was validated by assessing the fit of the items to a Rasch partial credit model (PCM) within the framework of item response theory (IRT). The PCM is optimal for polytomous data, such as that obtained from the 5-point Likert scales used in this study (Makransky, Lilleholt, & Aalby, 2017). The PCM is preferable to two or three parameter IRT models in settings where summed, total scores are used (Makransky & Bilenberg, 2014), because data are tested against the assumptions of the model, and, if met, the raw score of a scale is statistically sufficient (Tennant & Connaghan, 2007). In other words, a person's total score contains all of the information available within the specified context, which is a property that is unique to Rasch family models including the PCM (Makransky, Creed, & Rogers, 2014), meaning that the PCM is optimal for the UiL because total scale scores can be used as estimates of the latent trait.

Analyses were conducted using the Rasch Unidimensional Measurement Models program (RUMM2030; Andrich, Sheridan, & Luo, 2010), which is one of the most commonly used programs for Rasch analysis. The evaluation criteria that were applied included the assessment of unidimensionality, local dependence, item fit, measurement invariance in the form of differential item functioning (DIF) by grade level, and reliability as evaluated by the Person Separation Index (PSI) and Cronbach's alpha. The evaluation criteria have been described elsewhere and are only reviewed here briefly (for more information, see Christensen et al., 2016; Nielsen et al., 2017a,b; Pallant & Tennant, 2007; Tennant & Conaghan, 2007).

Unidimensionality was evaluated according to a formal test proposed by Smith (2002). This test uses the first residual factor in a principal components analysis (of residuals) to determine two groups of items: those with positive and those with negative residuals. Each set of items is then used to calculate an independent trait estimate for each person in the sample. When items form a unidimensional scale, it is expected that the person estimates from the two item subsets should be similar. An independent samples *t*-test is used to determine whether there is a significant difference between the two person estimates. This is repeated for each person with the expectation that the percentage of tests lying outside the range of $\pm$1.96 should not exceed 5% (Makransky & Bilenberg, 2014).

Local dependence is an indicator of redundancy among items in a scale. This was assessed by investigating if residual correlations among the items in each scale were larger than the critical value, which was calculated based on a parametric bootstrapping of the Q_3 statistic, as described in Christensen et al. (2016).

Measurement invariance in the form of DIF across grades for each of the items in the UiL was also investigated. Items with significant Chi-square statistics at the .05 level (2-sided and with a Bonferroni correction applied separately within each DIF variable) are reported as exhibiting DIF.

Item fit was investigated in order to determine if all items were equally important for measuring a latent trait. Over-fit is obtained when an item discriminates (between individuals who are low and high on the non-cognitive trait) more than is expected by the model. Under-fit is obtained when the item does not discriminate as well as expected, which is an indication that the item measures something other than what was intended. Item fit was assessed by taking a random sub-sample of 400 students for each scale, since previous studies have found that the item fit statistics in RUMM can be biased with large sample sizes (400 is optimal for assessing item fit; Bergh, 2015; Hagell, & Westergren, 2016; Müller & Kreiner, 2015). Items were identified as not fitting the model when they had a fit residual > $\pm$2.5.

Reliability was assessed using Cronbach's alpha and the Person Separation Index (PSI). Although similar interpretations can be made with the two measures, the PSI takes the targeting of the scale into account because it is based on the IRT perspective that the standard error of a scale can vary at different points of the latent trait depending on the item information that is available at that point (Tennant & Cohaghan, 2007).

Table 1 summarizes the results and the changes that were made based on these analyses. A good fit was found for five scales, which were retained in their original form. Items did not function optimally in 11 scales; thus, items were either revised or re-written based on the results. Finally, the unidimensionality test showed that the Engagement Scale was multidimensional. This is consistent with Fredricks et al. (2005), and is not surprising as the scale had been formed with items from several subscales including behavioural, emotional, and academic engagement. Therefore, new items were written to cover all three of the engagement subscales. This resulted in a total of 19 scales in the UiL. These changes were then validated in a subsequent study in the same eight schools.

Validation study

This study aimed to validate the 19 scales (115 items) in the UiL using CFA and PCM. CFA is optimal for testing the construct validity of a test battery by identifying if the general structure of the battery fits an a-priori model (Fabrigar, Wegener, MacCallum, & Strahan, 1999). We investigated convergent and discriminant validity by testing if the data fit a 19-factor model. That is, the items in the 19 scales included in the UiL should measure the construct they are intended to measure, and not another construct. This is important because each scale in the test battery should provide unique information and not overlap with another scale. Once a general structure is identified, the PCM (within IRT) is optimal for investigating the quality of each item in each scale (Hays, Morales, & Reise, 2000) because it provides detailed information about the validity of a measurement instrument based on whether the instrument lives up to the set of assumptions described above.

Method

Participants and Procedure

The sample consisted of 1373 students (48.6% boys, ages 9 to 16 years) from 4th (N = 206), 5th (N = 200), 6th (N = 249), 7th (N = 192), 8th (N = 302), and 9th (N = 224) grades, who were assessed between January and March, 2017. The same classes from the eight schools used in the development phase were re-surveyed using the same data collection procedures across all schools. All teachers from 4th to 9th grade were invited to participate, and students were asked to respond to all 19 scales (115 items). As previously, the data were collected in two waves due to the large number of items (86% responded to all items in both waves). Student responses were matched using student ID. Student responses were discarded when there was a problem with the student ID or when a student responded to fewer than 10 items (this resulted in approximately 3% of the responses being discarded).

Measures

See Appendix 2 for the 115 items used to measure the 19 non-cognitive skills scales included in the final version of the UiL.

---Insert Table 2 here---

Analytic Strategy

To investigate the dimensionality of the UiL, CFA analyses were conducted in Mplus V7 (Muthen & Muthen, 2012) using polychoric correlations. An a-priori model with 19 factors that defined the items representing the 19 scales was tested. Reported goodness-of-fit indices included the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA). An acceptable fit is indicated by CFI and TLI ≥ .90, and RMSEA ≤ .06 (Hu & Bentler, 1999). The estimation method used was Muthen's (1984) 3-step procedure. Additionally, each of the 19 scales was validated by assessing the fit of the items to the PCM (Masters, 1982). These analyses were conducted with RUMM2030 (Andrich et al., 2010) using the same evaluation criteria as in the development phase (assessment of unidimensionality, local dependence, item fit, DIF by grade, and reliability as evaluated by PSI and Cronbach's alpha).

Results

Confirmatory Factor Analysis

The results of the CFA showed a good fit to the 19 factor model (CFI = .90; TLI = .90; RMSEA = .04). All items with the exception of Item 5 from the extrinsic motivation scale (“I only really do my best if I get rewarded”), loaded onto the intended scale as expected (see CFA item loadings in Appendix 2). The results support the general construct validity of the UiL, with the CFA providing evidence of convergent and discriminant validity as all but one of the items in the test battery measured the latent trait that was intended.

Partial Credit Model

Although the results of the CFA supported the general construct validity of the UiL, the PCM can provide more detailed information about how the items in each scale function as unidimensional sufficient scales (Makransky, Lilleholt, & Aaby, 2017). Table 3 reports the results of the PCM analyses. The item fit statistics are also presented in Appendix 2.

---Insert Table 3 here---

Intrinsic motivation: The six items had acceptable fit to the model. Reliability was acceptable with values of .81 and .76 for Cronbach's alpha and PSI, respectively. The unidimensionality test was slightly over the criteria of 5% (5.03%). There was no local dependence detected between any of the item pairs. Finally, uniform DIF was detected across grade levels for Item 3 ("I like activities that help me learn a lot"), which suggested that students in lower grades tended to have higher scores on the item compared to students in higher grades despite the same trait level.

Self-efficacy: Reliability was acceptable with values of .89 and .86 for alpha and PSI, respectively. No local dependence was detected. One of the six items had a negative fit residual over the acceptable criteria (Item 3: "I am sure that I can learn the material I am taught"), indicating that this item discriminated better than expected between students who were low and those who were high on the trait. The unidimensionality test was also marginally over the criterion of 5% (5.24%). Finally, uniform DIF was detected across grade levels for Items 4 ("I believe that I will get very good results in tests") and 6 ("I'm sure I can get better grades/evaluation than most other students"). The DIF pattern suggested that students in lower grades tended to have lower scores on Item 4 and higher scores on Item 6 compared to older children at the same trait level.

Self-regulation: All eight items had acceptable fit. Alpha was .69 and PSI was .70. There was local dependence indicating redundancy among three items: Items 2 ("I often say inappropriate things"), 6 ("I am not so good at concentrating"), and 7 ("I sometimes do things without

consideration"). The unidimensionality test was over criterion (8.38% significant tests). However, there was no DIF by grade.

Perseverance: All six items had acceptable fit. Reliability was acceptable with values of .79 and .73 for alpha and PSI, respectively. The unidimensionality test was acceptable, and there were no items exhibiting local dependence. Two items exhibited uniform DIF by grade: Item 4 ("I have defied resistance to solve important tasks") and Item 6 ("I keep trying even though I am about to lose"). Students in lower grades tended to have lower scores on Item 4 and higher scores on Item 6 compared to older children at the same trait level.

Conscientiousness: All six items had acceptable fit. Reliability was acceptable with values of .78 and .73 for alpha and PSI, respectively. The unidimensionality test was also acceptable, and there were no items exhibiting local dependence or DIF by grade.

Engagement: The items in each of the three engagement scales had acceptable fit, with the exception of the emotional engagement Item 1 ("I like being in school"). Reliability was acceptable with alpha values ranging from .81 to .87 and PSIs from .76 to .82. The unidimensionality test was marginally over criterion for behavioural (5.46%) and emotional engagement (5.46%); however, there were no items with local dependence. Finally, the behavioural engagement Item 4 ("I never make noise in class") and the academic engagement Items 1 ("I often read my school books, even though I do not have homework") and 5 ("I seek knowledge about what we learn in school through internet, TV, other books)" exhibited uniform DIF by grade. Students in lower grades tended to have score higher on Items 4 and 5, but lower on Item 1 compared to older children at the same trait level.

Cooperation: All five items had acceptable fit. Reliability was acceptable with values of .83 and .75 for alpha and PSI, respectively. The unidimensionality test was also acceptable, and there were no items exhibiting local dependence or DIF by grade.

Resilience: All four items had acceptable fit. Reliability was acceptable with values of .77 and .72 for alpha and PSI, respectively. The unidimensionality test was acceptable, and there were no items exhibiting local dependence. Finally, Items 1 (“I do not get stressed because of school work”) and 4 (“I'm good at dealing with adversity, e.g., bad evaluations, negative feedback”) exhibited uniform DIF by grade, but the pattern and direction were not clear.

Attention: All six items had acceptable fit. Reliability was .81 and .79 for alpha and PSI, respectively. The unidimensionality test was also slightly over criterion (5.17%), but no item exhibited local dependence. Finally, Item 4 (“I can easily work with an interesting assignment for several consecutive hours without a break”) exhibited uniform DIF by grade, with students in lower grades tending to have higher scores compared to older children at the same trait level.

Extrinsic motivation: With the exception of Item 5 (“I only really do my best if I get rewarded”), all items had acceptable fit. One potential explanation for this lack of fit is that the wording of the item in Danish may be ambiguous (see Appendix). Reliability was .77 and .75 for alpha and PSI, respectively. The unidimensionality test was also acceptable, and there were no items exhibiting local dependence. Item 1 (“What I prefer in school is to get praise from the teacher”) and Item 5 exhibited uniform DIF by grade. Students in lower grades tended to have lower scores on Item 1 and higher scores on Item 5 compared to older children at the same level.

Drive: All five items had acceptable fit. Reliability was .79 (alpha) and .78 (PSI). The unidimensionality test was marginally over criterion (5.17%), but no item exhibited local dependence. Item 1 (“When I believe in an idea, nothing can prevent me from implementing it”) exhibited non-uniform DIF by grade.

Critical thinking: All five items had acceptable fit. Reliability was also acceptable with .73 for alpha and .70 for PSI. Unidimensionality was acceptable, and there were no items exhibiting local dependence or DIF by grade.

Creativity: All five items had acceptable fit and reliability (.84 for alpha and .81 for PSI). Unidimensionality was satisfactory, and no items exhibited local dependence or DIF by grade.

Well-being: One of the seven items had a negative fit residual and did not fit the PCM (Item 5: "I am usually in a good mood"). Reliability was acceptable: with alpha at .88 and PSI at .73. Unidimensionality was confirmed, and there were no items exhibiting local dependence. Uniform DIF by grade was exhibited by Item 5 and Item 6 ("I get along with others"), with students in lower grades tending to have lower scores for Item 5 and higher scores for Item 6 compared to older children at the same trait level.

Self-esteem: One of the six items did not fit the PCM (Item 4: "I am very proud of who I am"). Reliability was good, with alpha of .91 and PSI of .83. Unidimensionality was acceptable, and no item exhibited local dependence. Uniform DIF by grade was exhibited by Item 1 ("I am generally pleased with myself") and Item 6 ("I have a lot to contribute"); however, the pattern and direction were not clear.

Empathy: Two of the seven items did not fit the PCM. Item 2 ("I am often affected by my friends' mood") had a positive fit residual indicating that the item did not discriminate as highly between students who were low and high on the trait. Item 6 ("I am really affected by how other people are doing") had a negative fit residual indicating that the item discriminated more than expected. Also, Item 7 ("I get happy when I see others who are happy") exhibited uniform DIF by grade, but there was not a clear pattern of direction. Reliability (alpha = .82, PSI = .75) and unidimensionality were acceptable, and no item exhibited local dependence.

Outcome expectations: Four of the 12 items did not fit the PCM. Item 12 ("Friends and family will be happy if I'm good at school") had a positive fit residual, and Items 5 ("By going to school, I can get good friends"), 7 ("By going to school, I can be good at helping others"), and 9 ("By going to school, I can be satisfied with myself") had negative fit residuals. Four item pairs

exhibited local dependence (Item 1, “If I do well in school, I believe I can get a well-paid job” with Item 3, “If I do well in school, I can get into the education of my choice; Item 1 with Item 4, “If I do well in school, I can complete an education of my choice”; Item 4 with Item 5; and Item 5 with Item 6, “By going to school, I can get good at working with others”). Reliability was sound with alpha of .90 and PSI of .79, and unidimensionality was acceptable. Finally, three items exhibited DIF by grade, including Item 2 (“If I do well in school, I believe I can get an exciting job”), Item 6, and Item 8 (“By going to school, I can improve my skills and become good at many things”). Students in lower grades tended to have higher scores on Item 2, but lower scores on Item 8 compared to older children at the same trait level; the pattern for Item 6 was unclear.

Discussion

The early development of non-cognitive skills is an important factor in children’s education, later employment, and success in life (Almlund, 2011; Levin, 2013). There is considerable literature developing tests to measure specific, non-cognitive skills like intrinsic motivation, resilience, self-esteem, and empathy (Pellegrino, 2012). However, most of this research has been conducted in Anglo-American countries (Brunello, 2011). In the Nordic countries, non-cognitive skills have been important for centuries (Wandall, 2013), but there has not been a strong tradition for developing and validating scales to measure these non-cognitive skills in this part of the world. Further, most previous validation studies have focused on single scales (e.g., Ambiel et al., 2015; Duckworth & Quinn, 2009; Stump, 2010). This paper reported the results of an ambitious development and validation process. The UiL provides a comprehensive battery of scales for the assessment of a broad range of non-cognitive skills relevant in the Scandinavian context.

The results of the CFA indicated that the student data fit the hypothesized 19 factor model, which provided evidence for discriminant and convergent validity of the UiL. Only one item from the extrinsic motivation scale did not load sufficiently on the hypothesized factor as intended.

The results from the PCM indicated that all reliability estimates (i.e., Cronbach's alpha and PSI), were good for such short scales, with values ranging between .69 and .91 for alpha and .70 and .86 for the PSI. The PSI can be evaluated similarly to alpha, but tends to be lower when the items do not appropriately target the sample. The finding that the values for the two statistics were similar across all scales is an indication that the UiL scales are well targeted for 4th to 9th graders in Denmark.

The results from the PCM also indicated that the conscientiousness, cooperation, critical thinking, and creativity scales fit the model well. These results provide evidence to support the validity of these scales within a Danish primary and secondary school context, and support the large base of literature that has investigated the validity of these scales in different contexts (John & Srivastava, 2001; Orchard et al., 2012; Pintrich et al., 1991). The perseverance, academic engagement, and resilience scales fit the model within each grade, but these scales had items that exhibited DIF between grades. DIF across grades was identified in 22 of the 115 items (in 13 of the 19 scales). Taken as a whole, these results suggest that the UiL can be interpreted validly within grades, but caution should be exercised when the results are compared across grades. With the exception of one item from the drive scale, the DIF was uniform, which means that the conditional dependency is relatively invariant across the latent trait continuum. When DIF is uniform it can be corrected by using group specific item parameters (e.g., see Makransky & Glas, 2013; Makransky et al., 2014, 2015; Schnohr et al., 2013), which can resolve the issue when comparing across grade levels.

Another source of misfit to the PCM was that the unidimensionality test indicated that *t*-tests of 7 of the 19 scales exceeded 5%. Future research should investigate this issue since the magnitude of the misfit was very small, with the exception of the self-regulation scale, where 8.38% of the tests were significant. The problem with multidimensionality in the self-regulation

scale could be a consequence of some of the items being negatively worded (Barnette, 2000; Wang et al., 2015), as these were the only items in the UiL that were worded in this way. This suggests that the multidimensionality could be a result of response bias, but more research is needed before this scale can be used with confidence.

Evidence of local dependence (LD), which indicates redundancy, was only found among three items in the self-regulation scale, and four pairs of items in the outcome expectations scale. Although LD can be accounted for in the Rasch model by combining items into test-lets (a group of items related to a single content area that is developed as a unit; Christensen, Kreiner, & Mesbah, 2013), the finding for the outcome expectations scale suggests multidimensionality and future validation studies are needed before this scale can be used as a unidimensional measure. The item fit test also showed that there were 11 items (from 7 scales) that did not fit the PCM. Future validation efforts are needed before these items can be confidently used.

Practical implications and future research

One important practical implication when using the UiL is the limitation of using a self-report instrument. Non-cognitive data collected using self-report instruments can have reduced validity due to intentional or non-intentional response biases. These challenges need to be considered when using the UiL. Significant effort was spent ensuring that students and teachers were well informed and instructed about the purpose and context of the UiL, with the intention of creating trust that the data would be used for individual development purposes, in order to motivate precise and accurate responding. It is also important that teachers are confident that the results are collected for formative and pedagogical use and not for accountability. Further, the purpose of the UiL is to use the data for feedback in a dialogue with students in order to develop their life skills. The intention is to use the UiL regularly in order to track the development of individual students, which means that additional longitudinal studies need to be conducted to investigate the

measurement invariance of the specific items and scales over time. Related to this, more studies are needed to investigate the consequences of repeatedly responding to the UiL. Students who have completed the items several times, and have had feedback/discussions with their teacher, might respond differently to students who are taking the UiL for the first time. Some sources of misfit were identified in this study, which will require further evaluation in future studies. Future research – especially using longitudinal data – should re-evaluate the usefulness of the items that did not fit the PCM.

Additionally, it should be noted that the UiL does not need to be used as a complete battery. It provides a number of independent scales of specific non-cognitive skills that can be selected based on what is relevant for a given purpose. For example, if a school has an intervention in place to develop motivation, then the motivational scales can be used before and after the intervention to evaluate its effect.

The results also indicated that there was a lack of measurement invariance in the form of DIF across grade level for most of the scales. Future research should investigate the practical implications of the DIF identified in this study, as sample size was very large (which means that even small violations of measurement invariance become significant). Although group specific item parameters can be used when making comparisons across grades with the UiL (e.g., Makransky & Glas, 2013; Makransky et al., 2014; Schnohr et al., 2013), more research is needed to investigate the structure of constructs across this developmental range.

Finally, the two studies reported in this paper provide initial evidence for the validity and reliability of the UiL as a tool to assess 19 non-cognitive skills in school children in Denmark. This provides the first, validated, non-cognitive skills battery that can be applied by schools to follow the development of these skills throughout the education system. The next step is to assess the validity of using the UiL longitudinally. One of the practical applications of the UiL is to take

repeated measures of the non-cognitive skills of students during their education. This will make it possible to track children's non-cognitive skill development and investigate the antecedents and outcomes of these skills. Levels of non-cognitive skills can be correlated with data held in the Danish registry on the health, education, and economic status of citizens. Thus, the UiL is potentially an important tool that can facilitate linking children's development in non-cognitive skills and the long-term impact of these skills on health, education, economic, and life outcomes.

References

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. (2011). *Personality psychology and economics* (Working Paper 16822). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16822.pdf

Ambiel, R. A., Noronha, A. P. P., & de Francisco Carvalho, L. (2015). Analysis of the professional choice self-efficacy scale using the Rasch-Andrich rating scale model. *International Journal for Educational and Vocational Guidance*, *15*, 205-219.

Andrich, D., Sheridan, B., & Luo, G. (2010). *Rasch models for measurement*: RUMM2030. Perth, Australia: RUMM Laboratory.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency. *Education and Psychological Measurement*, 60, 361-370. http://dx.doi.org/10.1177/00131640021970592.

Barrick, M. R., & Mount, M. K. (1993). Autonomy as a moderator of the relationships between the Big Five personality dimensions and job performance. *Journal of Applied Psychology*, *78*, 111-117.

Bergh, D. (2015). Chi-squared test of fit and sample size: A comparison between a random sample approach and a chi-square value adjustment method. *Journal of Applied Measurement, 16*, 204–217.

Bonell, C., Humphrey, N., Fletcher, A., Moore, L., Anderson, R., & Campbell, R. (2014). Why schools should promote students' health and wellbeing. *British Medical Journal, 348,* 1–2.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Brunello, G., & Schlotter, M. (2011). Non-cognitive skills and personality traits: Labour market relevance and their development in education and training systems. Discussion paper No. 5743. Retrieved from https://www.econstor.eu/bitstream/10419/51586/1/669379840.pdf

Chatterji, M. (2013). *Validity and test use: An international dialogue on educational assessment, accountability and equity*. Bingley, UK: Emerald.

Christensen, K. B., Kreiner, S., & Mesbah, M. (Eds.). (2013). *Rasch models in health*. London, UK: John Wiley & Sons.

Christensen, K. B., Makransky, G., & Horton, M. (2016). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*, 178–194.

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology*, *111*, 225–236.

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin*, *125*, 276-288.

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*, 1087–1101.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*, 166-174.

Duckworth, A., & Seligman, M. (2005) Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*, 939–944.

Eignor, D. R. (2013). *The standards for educational and psychological testing*. American Psychological Association.

Embertson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Fredricks, J. A., Blumenfeld, P., Friedel, J., & Paris, A. (2005). School engagement. In K.A. Moore & L. Lippman (Eds.), *What do children need to flourish?* (pp. 305–321). New York, NY: Springer.

García, E. (2016). The need to address non-cognitive skills in the education policy agenda. In M. S. Khine & S. Areepattamannil (Eds.), *Non-cognitive skills and factors in educational attainment* (pp. 31–64). Rotterdam, ND: Sense Publishers.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcome measurement in the 21st century. *Medical Care, 38*, 1128-1142.

Hagell, P., & Westergren, A. (2016). Sample size and statistical conclusions from tests of fit to the Rasch Model according to the Rasch Unidimensional Measurement Model (RUMM) Program in health outcome measurement. *Journal of Applied Measurement*, *17*, 416–431.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Hove, UK: Psychology Press.

Hattie, J. A. (2013). Synlig læring - for lærere ”Visible learning – for teachers”. Frederikshavn: Dafolo.

Heckman, J. J. (2000). Policies to foster human capital. *Research in Economics*, *54*, 3–56.

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic Inquiry*, *46*, 289–324.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour economics*, *19*(4), 451-464.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., & Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative economics*, *1*(1), 1-46.

Hirschi, T., & Gottfredson, M. R. (1995). Control theory and the life-course perspective. *Studies on Crime & Crime Prevention, 4,* 131-142.

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*, 1–55.

Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, *65*, 19–51.

John, O. P., & Srivastava, S. (2001). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102 –138). New York, NY: The Guilford Press.

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the Basic Empathy Scale. *Journal of Adolescence*, *29*, 589–611.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. Discussion Paper, No. 8696. Retrieved from https://www.econstor.eu/bitstream/10419/107477/1/dp8696.pdf

Levin, H. M. (2012). More than just test scores. *Prospects*, *42*, 269–284.

Levin, H. M. (2013). The utility and need for incorporating noncognitive skills into large-scale educational assessments. In M. Von Davier, E. Gonzalez, I, Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 67–86). Dordrecht, ND: Springer Netherlands.

Liddle, I., & Carter, G. F. (2015). Emotional and psychological well-being in children: The development and validation of the Stirling Children's Well-being Scale. *Educational Psychology in Practice*, *31*, 174–185.

Makransky, G., & Bilenberg, N. (2014). Psychometric properties of the parent and teacher ADHD Rating Scale (ADHD-RS): Measurement invariance across gender, age and informant. *Assessment, 21*, 694–705.

Makransky, G., & Glas, C. A. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement*, *46*, 3228–3237.

Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, *72*, 276–285.

Makransky, G., Rogers, M. E., & Creed, P. A. (2015). Analysis of the construct validity and measurement invariance of the career decision self-efficacy scale: A Rasch model approach. *Journal of Career Assessment, 23*, 645–660.

Makransky, G., Schnohr, C., Torsheim, T., & Currie, C. (2014). Equating the HBSC Family Affluence Scale across survey years: A method to account for item parameter drift using the Rasch model. *Quality of Life Research, 23*, 2899–2907.

Martin, A. J., & Marsh, H. W. (2008). Academic buoyancy: Towards an understanding of students' everyday academic resilience. *Journal of School Psychology*, *46*, 53–83.

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

Mortimore, P., David-Evans, M., Laukkanen, R., & Valijarvi, J. (2004). OECD-rapport om grundskolen i Danmark "OECD-report on elementary school in Denmark". Retrieved from http://static.uvm.dk/publikationer/2004/oecd/hel.html

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115–132.

Muthen, L. K., & Muthen, B. O. (2012). Mplus (Version 7). Los Angeles, CA: USA.

Müller, M., & Kreiner, S. (2015). Item fit statistics in common software for Rasch analysis: Research report 15/06. Department of Biostatistics, University of Copenhagen. Retrieved from https://ifsv.sund.ku.dk/biostat/annualreport/images/2/2f/Research_Report_15-06.pdf

Nielsen, T., Vang, M. L., Dammeyer, J., Makransky, G. (2017). Gender fairness in self-efficacy? A Rasch-based validity study of the General Academic Self-efficacy scale (GASE). *Scandinavian Journal of Educational Research*. doi:10.1080/00313831.2017.1306796

Nielsen, T., Makransky, G., Vang, M. L., Dammeyer, J. (2017). How specific is specific self-efficacy? A construct validity study using Rasch measurement models. *Studies in Educational Evaluation*. doi:10.1016/j.stueduc.2017.04.003

Orchard, C. A., King, G. A., Khalili, H., & Bezzina, M. B. (2012). Assessment of Interprofessional Team Collaboration Scale (AITCS): Development and testing of the instrument. *Journal of Continuing Education in the Health Professions*, *32*, 58–67.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*, 1–18.

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

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire* (Technical Report 91-B-

004). National Center for Research to Improve Postsecondary Teaching and Learning. Retrieved from http://files.eric.ed.gov/fulltext/ED338122.pdf

Porath, C. L., & Bateman, T. S. (2006). Self-regulation: From goal orientation to job performance. *Journal of Applied Psychology*, *91*, 185–192.

Roberts, R.D., Martin, J.E., & Olaru, G. (2015). *A Rosetta Stone for noncognitive skills*. Retrieved from http://asiasociety.org/files/A_Rosetta_Stone_for_Noncognitive_Skills.pdf

Rosen, J. A., Glennie, E. J., Dalton, B. W., Lennie, J. M., & Bozick R. N. (2010). *Noncognitive skills in the classroom: New perspectives on educational research.* RTI International. Retrieved from http://www.rti.org/rtipress

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Schnohr, C. W., Makransky, G., Kreiner, S., Torsheim, T., Hofmann, F., De Clercq, B., ... & Currie, C. (2013). Item response drift in the Family Affluence Scale: A study on three consecutive surveys of the Health Behavior in School-aged Children (HBSC) survey. *Measurement, 46*, 3119–3126.

Stump, G. S. (2010). Development of a nursing student self-efficacy scale using item response theory. *Nursing Research, 61,* 149- 158.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care and Research, 57*, 1358–1362.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*, 271–324.

Wandall, J. (2013). Education, testing, and validity: A Nordic comparative perspective. In M. Chatterji (Ed.). *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 137–161). Bingley, UK: Emerald Group Publishing.

Wang, W. C., Chen, H. F., & Jin, K. Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, *75*(1), 157-178.

Table 1

Scales, Sources, and Number of Items included in the Quantitative Study

Scale	Source reference[1]	Initial number of items	Changes made after quantitative analyses and pilot test
1. Intrinsic motivation	Pintrich et al. (1991)	6	No change
2. Self-efficacy	Pintrich et al. (1991)	6	No change
3. Self-regulation	Tangney et al. (2004)	8	1 deleted, 1 new
4. Perseverance	Duckworth & Quinn (2009)	6	No change
5. Conscientiousness	John & Srivastava (2001)	7	2 deleted, 1 new
6. Cooperation	Orchard et al. (2012)	6	1 deleted
7. Resilience	Martin & Marsh (2008)	4	No change
8. Attention	Derryberry & Reed (2002)	9	3 deleted
9. Extrinsic motivation	Pintrich et al. (1991)	5	1 deleted, 2 new
10. Proactive behaviour/drive	Porath & Bateman (2006)	5	No change
11. Critical thinking	Pintrich et al. (1991)	6	1 deleted
12. Creativity/openness	John & Srivastava (2001)	6	2 deleted, 1 new
13. Engagement	Fredricks et al. (2005)	7	Split into 3 subscales
14. Well-being	Liddle & Carter (2015)	8	1 deleted
15. Self-esteem	Rosenberg (1965)	7	5 deleted, 4 new
16. Outcome expectations	Lent et al. (1994)	12	1 deleted, 1 new
17. Empathy	Jolliffe & Farrington (2006)	10	6 deleted, 3 new

Table 2

Scales, Number of Items, and Example Item included in the Final UiL.

Scale	# of items	Example item (Danish translation into English)
1. Intrinsic motivation	6	"I like activities where I learn a lot"
2. Self-efficacy	6	"I think I will do well in school"
3. Self-regulation	8	"I am good at resisting temptation"
4. Perseverance	6	"I keep on fighting, even when success is not likely"
5. Conscientiousness	6	"I am thorough"
6. Behavioral engagement	5	"I never make trouble at school"
7. Cognitive engagement	5	"I often read my schoolbooks, even though I don't have homework"
8. Cooperation	5	" I am always open and honest to others"
9. Resilience	4	"I don't let bad grades/evaluations affect me"
10. Attention	6	"I can easily work at an interesting assignment for several consecutive hours without a break"
11. Extrinsic motivation	6	"I want to do well in school so friends and family can see that I am successful"
12. Proactive behaviour/drive	5	"When I can think of a better way to do things, I try and do it that way"
13. Critical thinking	5	"I always evaluate if what I do in school is meaningful"
14. Creativity/openness	5	"I am curious about a lot of things"
15. Emotional engagement	5	"I like being at school"
16. Well-being	7	"I am usually in a good mood"
17. Self-esteem	6	"I am very proud of who I am"
18. Outcome expectations	12	"If I do well in school, I think I can get an exciting job"
19. Empathy	7	"When I see happy people, I become happy myself"

Table 3

Results of the PCM Analysis for the 19 Scales in the Final UiL.

Scale	Reliability		Fit			
	Alpha	PSI	Unidim.	LD	Item fit	DIF
Intrinsic motivation	.81	.76	5.03%	OK	OK	3
Self-efficacy	.89	.86	5.24%	OK	5 (-)	4,6
Self-regulation	.69	.70	8.38%	3 LD	OK	OK
Perseverance	.79	.73	3.06%	OK	OK	4,6
Conscientiousness	.78	.75	4.01%	OK	OK	OK
Engagement-Behavioral	.81	.76	5.46%	OK	OK	4
Engagement-Academic	.82	.80	4.88%	OK	OK	1,5
Engagement-Emotional	.87	.82	5.46%	OK	1 (-)	OK
Cooperation	.83	.75	3.50%	OK	OK	OK
Resilience	.77	.72	4.95%	OK	OK	1,4
Attention	.81	.79	5.17%	OK	OK	4
Extrinsic motivation	.77	.75	4.66%	OK	5 (+)	1,5
Drive	.79	.78	5.17%	OK	OK	1
Critical thinking	.73	.70	2.77%	OK	OK	OK
Creativity	.84	.81	4.15%	OK	OK	OK
Well-being	.88	.73	1.24%	OK	5 (-)	5,6
Self-esteem	.91	.83	4.44%	OK	4 (-)	1,6
Empathy	.82	.75	3.93%	OK	2 (+), 6 (-)	7
Outcome expectations	.90	.79	4.95%	4 LD	5 (-), 7(-), 9 (-), 12(+)	2,6,8

Alpha = Cronbach's alpha reliability; PSI = Person Separation Index; Unidim = unidimensionality test; LD = local dependence; DIF = differential item functioning.