Development and Validation of the Development in Life Skills (UiL): A Battery of Non-cognitive skills for Danish School Children

Abstract

Abundant research from the last decades has pointed at the importance of non-cognitive skills for success in life and at the malleability of these factors. In Nordic countries, this has had a strong impact on the policy focus on assessing and implementing interventions to increase non-cognitive skills such as self-control, resilience, intrinsic motivation, and self-efficacy in school-age children. This paper describes the development and validation of a Danish measurement tool called "UiL - Udvikling i Livsfærdigheder" (which means Development in Life Skills) designed to measure 19 non-cognitive constructs that have been identified as being important for the development of primary and lower secondary school children. Two studies were conducted. The first study describes the development of the UiL and the preliminary validation with 1560 students. The second study describes validation of the final UiL (which included 115 items from 19 scales). Confirmatory factor analysis (CFA) and the partial credit model (PCM) were used to investigate the psychometric quality of the scales. There was overall good model-data fit of the scales to the a-priori CFA model. Each of the 19 scales also had acceptable fit to the PCM; however, some sources of misfit were identified suggesting small revisions.

Introduction

In most countries, the primary purpose of basic school education is to develop cognitive knowledge and skills in subjects like reading, maths, science, and history (e.g., Bonell, 2014; Garcia 2016; Heckmann, 2008; Levin, 2012; Pellegrino, 2012). However, in many countries, education systems also have a charter to develop non-cognitive skills and attitudes. For example, the Danish Act of Primary and Lower Secondary School (Ministry of Justice, 2006) statement of purpose emphasizes how schools, through teaching in academic subjects, should also promote students' personal and social development (e.g., motivation, awareness, imagination, and self-confidence) and develop students as responsible democratic citizens (Wandall, 2013). These non-cognitive skills, of which there are many including grit or perseverance, self-discipline or self-regulation, and self-efficacy (see Roberts, Martin, & Olaru, 2015; Rosen, Glennie, Dalton, Lennie, & Bozick, 2010), have been shown to be important predictors of long-term outcomes such as achieving higher levels of education, better job performance, lower criminality, and better general happiness (e.g., Barrick et al., 2001; Diener & Lucas, 1999; Duckworth & Seligman, 2005; Poropat, 2009). Further, research has pointed to the malleability of these skills (Kautz et al., 2014).

The Danish National School Reform (Undervisningsministeriet 2013) highlighted the need for valid assessment instruments that can assess a wide array of non-cognitive skills. Furthermore, inspired by Hattie's (Hattie 2013) feed-up, feed-back, and feed-forward model for making learning and learning goals visible for students and teachers, schools and stakeholders also are requested to use measurement tools to monitor students well-being and evaluate their interventions (Undervisningsministeriet 2017). With these needs in mind, the aim of the current research was to develop and validate a Danish battery to measure a wide range of non-cognitive skills that have been identified as being central to education in Denmark. This battery was called

UiL (*Udviklings i Livsfærdigheder* translated as *Development in Life Skills*). This paper describes the two studies to develop and then validate the UiL using samples of Grade 4 to 9 Danish pupils.

Study 1

The preliminary version of the UiL battery was developed by means of a comprehensive analysis process based on the American Educational Research Association (1999). 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

Qualitative Procedures and Participants

A panel of five international experts from different relevant areas (including psychology, psychometrics, teaching, and economy) worked together with Danish researchers, consultants, and employees (teachers and civil servants) from the two largest municipalities in Denmark (Copenhagen and Aarhus) to develop a list of non-cognitive constructs that would be measured in the UiL. The statement of purpose from the Act of Primary and Lower Secondary School (Ministry of Justice, 2006) as well as a literature review of relevant scales were used to identify an initial list of constructs. 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* i*n their own ability, take a stand*, and *take action*". Thus, constructs such as creativity, critical thinking, self-efficacy, self-esteem, and drive are relevant to measure. As a result of this process, the panel identified 17 constructs (see Table 1). A literature review was then conducted to find the most relevant existing literature on each of the constructs, along with the most suitable and psychometrically sound scales.

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 items needed to be made relevant for Danish students. Adaptation was required as some concepts had different connotations in Danish. For example: "Hard work" is something that most American parents' want their children to learn, admire, and respect. However, to most Danes, the concept of "Hard work" is not admirable – especially not for children where independence and tolerance are more admirable qualities (Inglehart, 2000). The items were then reviewed by the panel of experts and (a) proofread by civil servants and pedagogical consultants from the municipalities for content, item relevance, language, and spelling, (b) evaluated in a small pilot with twenty-five 5th grade students in one of the target schools, (c) revised by the research group following student feedback on word and sentence difficulty and length of survey, and (d) added to a questionnaire with a 5-point Likert scale that ranged from "*completely agree*" (1) to "*completely disagree*" (5).

----------Insert Table 1 here----------

Quantitative Procedures and Participants

The sample consisted of valid results from 1560 students (50.4% boys) between 4th and 9th Grades from 8 elementary schools in Denmark. Data collection took place from October 25th to November 18th 2016. Each of the 17 scales from Table 1 was validated by assessing the fit of the items to the partial credit Rasch model (PCM; Masters, 1982). Analyses were conducted with RUMM2030 (Andrich, Sheridan, & Luo, 2010). The evaluation criteria 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 Authors, 2016; Tennant & Conaghan, 2007; Pallant & Tennant, 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 2 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 -1.96 to 1.96 should not exceed 5% (Authors, 2014).

Local dependence (an indicator of redundancy among items in a scale) 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 0.05 level (2-sided and with a Bonferroni correction applied separately within each DIF-variable) are reported as exhibiting DIF. Finally, 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 are biased with large sample sizes (a sample size of 400 is optimal for assessing item fit in the program; Bergh, 2015; Hagell, & Westergren, 2016; Müller & Kreiner, 2015). Items were identified as not fitting the model when they had a fit residual over +/-3.

The final column of Table 1 summarizes the changes that were made based on these analyses. Acceptable fit was found for five scales, so these were retained in their original format. The items did not function optimally in 11 scales, so 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 sub-scales including behavioural, emotional, and cognitive engagement. Therefore, new items were written to cover all three of the engagement sub-scales. 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.

Study 2

Study 2 aimed to confirm the findings of Study 1 using CFA (Brown, 2015) and the PCM, and validate the revised items resulting from Study 1.

Method

Participants and Procedure

The sample consisted of 1373 students (48.6% boys) 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 25th and March 10th 2017 in the eight schools.

Measures

Table 2 provides an overview of the 19 non-cognitive skills scales comprising 115 items that were included in the final version of the UiL.

----------Insert Table 2 here----------

Analytic Strategy

Missing data were excluded from the list. CFA analysis to investigate the dimensionality of the UiL was conducted in Mplus Version 7 (Muthen & Muthen, 2012) using polychoric correlations. Reported goodness-of-fit indices include 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 $\geq$ 0.90, and RMSEA $\leq$ 0.06 (Hu & Bentler, 1999). The estimation method used was Muthen's three-step procedure (Muthen, 1984). Additionally, each of the 19 scales was also

validated by assessing the fit of the items to the PCM (Masters, 1982) as in Study 1. Analyses were conducted with RUMM2030 (Andrich et al., 2010). The same evaluation criteria were used as in Study 1 (assessment of unidimensionality, local dependence, item fit, DIF by grade level, 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 model (RMSEA = 0.04; CFI = 0.90; TLI = 0.90). All items, with the exception of item 5 in the extrinsic motivation scale, loaded on the intended scale as expected (see Table 3). The results support the general construct validity of the UiL.

----------Insert Table 3 here----------

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 (Authors, 2017). Table 4 reports the results of the PCM analyses. The reliability estimates, in terms of Cronbach's alpha and the PSI, were acceptable with values ranging between .69 and .91 for alpha, and .70 and .86 for the PSI, respectively. 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 appropriately targeted for 4^{th} to 9^{th} graders in Denmark.

The unidimensionality test indicated that *t*-tests of 7 of the 19 scales exceeded 5%. However, the magnitude of the misfit was very small, with the exception of self-regulation, which had 8.38% significant tests. Therefore, there is only evidence of multidimentionality for the self-regulation scale. An inspection of this scale showed that the misfit was due to three of the items

being negatively worded, which suggests that the multidimensionality could be a result of response bias. Evidence of local dependence (LD), which indicates redundancy, was only found between 1 pair of items in the self-regulation scale, and 4 pairs of items in the outcome expectations scale.

The item fit test showed that all of the items fit the PCM with the exception of item 5 in the self-efficacy scale, item 6 in the empathy scale, and item 7 in the outcome expectations scale (which discriminated more than the other items in the respective scales). Furthermore, item 5 in the extrinsic motivation scale (this item was also identified as problematic in the CFA), item 2 in the empathy scale, and item 1 in the outcome expectations scale had poorer discrimination than the other items in the respective scales.

The final column of Table 4 shows the items that exhibited DIF by grade level. DIF by grade was identified in 12 of the 19 scales, which suggests that group specific item parameters should be used when comparing results across grade levels (for an example, see Authors, 2013; Authors, 2014; Schnohr et al., 2013).

----------Insert Table 4 here----------

Discussion

The early development of non-cognitive skills has been found to be an important factor for children's education, later employment, and success in life (Almlund 2011; Garcia 2013; Levin 2013). Although there is abundant literature investigating specific, non-cognitive skills like intrinsic motivation, resilience, self-esteem, and empathy (Pellegrino 2012), and numerous scales to measure specific, individual non-cognitive skills, the UiL provides a comprehensive battery of scales for the assessment of a broad range of the most important non-cognitive skills.

The results of the two studies conducted on a large sample of Danish primary to lower secondary students, indicated that the 19 scales in the UiL have good psychometric properties. The

CFA indicated that the structure of the UiL matched the a-priori hypothesized structure. The results of the PCM analysis indicated that the scales had good fit to the model in general. Some sources of misfit were identified, which will require further evaluation in future studies. Future studies are also needed to clarify the structure of the self-regulation scale, as there was evidence of potential multidimensionality. Further, there was some evidence of local dependency in this scale as well as in the outcome expectations scale. Future research should re-evaluate the usefulness of these items that did not fit the PCM.

The results indicated that there was a lack of measurement invariance in the form of DIF across grade level for most of the scales in the UiL. 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 (Authors, 2013; Authors, 2014; Schnohr et al., 2013), more research is needed to investigate the structure of constructs across this developmental range.

Finally, it should be noted that the UiL is not necessarily intended 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 pre- and post- to evaluate this.

In general, the two studies provide evidence of the validity and reliability of the UiL as a tool to measure 19 non-cognitive skills for 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 will be to assess the validity of using the UiL longitudinally, as one of the potential practical applications of UiL is to take repeated

measures of students' non-cognitive skills throughout their education. This will make it possible to track children's non-cognitive skill development and investigate the achievement and other outcomes of these skills. Levels of these non-cognitive skills will be able to be correlated with data held in the Danish registry on the health, education, and economic status of citizens. Therefore, the UiL provides an important tool to link children's development in non-cognitive skills and the long term impact of these skills on health, educational, economic, and life outcomes.

References

Authors. (2016)

Authors. (2013)

Authors. (2014)

Authors. (2015)

Authors. (2017)

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (pp.11–17). Washington, DC: American Educational Research Association.

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

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*(2), 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.* http://dx.doi.org.libraryproxy.griffith.edu.au/10.1136/bmj.g3078

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

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

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology, 111*(2), 225–236. doi:10.1037//0021-843X.111.2.225

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. doi:10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166–174. doi:10.1080/00223890802634290

Duckworth, A., & Seligman, M. (2005) Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*(12), 939–944. doi:10.1111/j.1467-9280.2005.01641.x

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

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: SensePublishers.

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*(4), 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. Frederikshavn: Dafolo.

Heckman, J. J. (2000). Policies to foster human capital. *Research in economics*, *54*(1), 3–56. https://doi.org/10.1006/reec.1999.0225

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic inquiry*, *46*(3), 289–324. https://doi.org/10.3386/w14064

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.

Inglehart, R., & Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, *65*(1), 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*(4), 589–611.

Levin, H. M. (2012). More than just test scores. *Prospects*, *42*(3), 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: Springer Netherlands.

Liddle, I., & Carter, G. F. A. (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*(2), 174–185. https://doi.org/10.1080/02667363.2015.1008409

Martin, A. J., & Marsh, H. W. (2008). Academic buoyancy: Towards an understanding of students' everyday academic resilience. *Journal of School Psychology*, *46*(1), 53–83. https://doi.org/10.1016/j.jsp.2007.01.002

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

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

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*(1), 58–67. https://doi.org/10.1002/chp.21123

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), 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*(1), 185–192.

Roberts, R.D., Martin, J.E., & Olaru, G. (2015). *A Rosetta Stone for noncognitive skills*. Available at 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, Available at http://www.rti.org/rtipress

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

Schnohr, C.,. . . 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*(9), 3119–3126. https://doi.org/10.1016/j.measurement.2013.06.016

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*(8), 1358–1362. https://doi.org/10.1002/art.23108

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*(2), 271–324.

Undervisningsministeriet. (2013). *Aftale mellem regeringen (Socialdemokraterne, Radikale Venstre og Socialistisk Folkeparti), Venstre og Dansk Folkeparti om et fagligt løft af folkeskolen.* Retrieved from https://uvm.dk/-/media/filer/uvm/udd/folke/pdf14/okt/141010-endelig-aftaletekst-7-6-2013.pdf?la=da

Undervisningsministeriet. (2017). *Trivselsmåling - Elevernes trivsel måles hvert år på alle landets folkeskoler.* Retrieved from https://uvm.dk/folkeskolen/elevplaner-nationale-test-og-trivselsmaaling/trivselsmaaling

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.

Table 1

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

Scale	Source reference[1]	Initial number of items	Changes made from quantitative analyses and the 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 behavior/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	Divided into 3 sub-scales
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 temptations"
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 existing job"
19. Empathy	7	"When I see happy people, I become happy myself"

Table 3

CFA Results including STDYX Standardization Estimates and p-values for all Items in the Final UiL.

Intrinsic motivation (IM)		Self-efficacy (SE)		Attention (AT)		Perseverance (PR)		Conscientiousness (CC)	
Item	Est.	Item	Est.	Item	Est.	Item	Est.	Item	Est.
IM_1	0.794	SE_1	0.875	AT_1	0.758	PR_1	0.647	CC_1	0.689
IM_2	0.743	SE_2	0.782	AT_2	0.612	PR_2	0.735	CC_2	0.423
IM_3	0.755	SE_3	0.816	AT_3	0.683	PR_3	0.690	CC_3	0.744
IM_4	0.637	SE_4	0.821	AT_4	0.625	PR_4	0.650	CC_4	0.770
IM_5	0.584	SE_5	0.922	AT_5	0.647	PR_5	0.666	CC_5	0.623
IM_6	0.734	SE_6	0.701	AT_6	0.717	PR_6	0.662	CC_6	0.777
Engagement behavioural (EB)		Engagement cognitive (EC)		Engagement emotional (EE)		Cooperation (CO)		Creativity (CR)	
EB_1	0.735	EC_1	0.717	EE_1	0.840	CO_1	0.677	CR_1	0.818
EB _2	0.704	EC_2	0.798	EE_2	0.877	CO_2	0.764	CR_2	0.772
EB _3	0.938	EC _3	0.786	EE_3	0.779	CO_3	0.795	CR_3	0.702
EB _4	0.656	EC _4	0.703	EE_4	0.696	CO_4	0.815	CR_4	0.798
EB _5	0.650	EC _5	0.682	EE_5	0.890	CO_5	0.780	CR_5	0.731
Self-esteem (ES)		Extrinsic Motivation (EM)		Critical Thinking (CT)		Drive (DR)		Resilience (RE)	
ES_1	0.855	EM_1	0.723	CT_1	0.842	DR_1	0.737	RE_1	0.626
ES_2	0.851	EM_2	0.562	CT_2	0.593	DR_2	0.698	RE_2	0.941
ES_3	0.918	EM_3	0.710	CT_3	0.444	DR_3	0.766	RE_3	0.713
ES_4	0.882	EM_4	0.850	CT_4	0.540	DR_4	0.621	RE_4	0.700
ES_5	0.826	EM_5	0.013	CT_5	0.699	DR_5	0.769		
ES_6	0.835	EM_6	0.644						
Self-regulation (SR)		Empathy (EP)		Well Being (WB)		Outcome Expectations (OE)			
SR_1	0.538	EP_1	0.621	WB_1	0.818	OE_1	0.827	OE_7	0.845
SR_2	0.263	EP_2	0.520	WB_2	0.629	OE_2	0.840	OE_8	0.861
SR_3	0.498	EP_3	0.912	WB_3	0.796	OE_3	0.825	OE_9	0.790
SR_4	0.556	EP_4	0.766	WB_4	0.791	OE_4	0.836	OE10	0.797
SR_5	0.764	EP_5	0.573	WB_5	0.808	OE_5	0.783	OE11	0.875
SR_6	0.538	EP_6	0.720	WB_6	0.791	OE_6	0.853	OE12	0.591
SR_7	0.271	EP_7	0.865	WB_7	0.823				
SR_8	0.514								

Note: all p values were < .001 with the exception of EM 5 which had a p value of 0.754.

Table 4

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%	1 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-Cognitive	.82	.80	4.88%	OK	OK	1,5
Engagement-Emotional	.87	.82	5.46%	OK	OK	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	OK	5,6
Self-esteem	.91	.83	4.44%	OK	OK	1,6
Empathy	.82	.75	3.93%	OK	2 (+), 6 (-)	7
Outcome expectations	.90	.79	4.95%	4 LD	7(-), 12(+)	2,6,8