Accepted Manuscript

Evaluating the Leuven Research Skills Test for 11th and 12th grade
Louise Maddens, Fien Depaepe, Rianne Janssen, Annelies Raes, Jan Elen

DOI: <https://doi.org/10.1177/0734282918825040>

To appear in: Journal of Psychoeducational Assessment

Accepted Date: 24 December 2018

Please cite this article as: Maddens, L., Depaepe, F., Janssen, R., Raes, A., & Elen, J. (2019). Evaluating the Leuven Research Skills Test for 11th and 12th grade. *Journal of Psychoeducational Assessment*. Advance online publication. doi: <https://doi.org/10.1177/0734282918825040>

This is a PDF file of an unedited manuscript that has been accepted for publication. The manuscript will undergo copyediting, typesetting and review of the resulting proof before it is published in its final form.

Evaluating the Leuven Research Skills Test for 11th and 12th grade

Louise Maddens^{ab}, Fien Depaepe^{ab}, Rianne Janssen^c, Annelies Raes^{ab}, Jan Elen^a

^aKU Leuven, Faculty of Psychology and Educational Sciences, Centre for Instructional Psychology and Technology

^bKU Leuven, ITEC, also at imec, Kapeldreef 75, B-3001 Leuven, Belgium

^cKU Leuven, Faculty of Psychology and Educational Sciences, Centre for Educational Effectiveness and Evaluation

Abstract: According to (inter)national policy and curriculum documents, the acquisition of research skills is an important objective of secondary education. However, the conceptualization and hence the operationalization of this concept seems ambiguous. Furthermore, no test exists to assess students' proficiency in (a broad range of) research skills in a 11th and 12th grade behavioral sciences classroom context. This paper first elaborates on what constitutes research skills in this educational context. Second, the development and testing process of the Leuven Research Skills Test (LRST) is described. Third, the psychometric properties and the dimensional structure of the LRST are presented, based on a large-scale sample ($n = 405$) of Belgian students in 11th and 12th grade. The results revealed that (1) the LRST is an internal consistent instrument and that (2) a hierarchical model with eight subordinate factors and one single uniting upper level factor appears to be the best fit to the data (in comparison with a unidimensional model, and an eight-factor multidimensional model). It is argued that the LRST can be used to assess (individual differences in) overall research skills proficiency and to investigate the effect of particular interventions to foster research skills in future studies.

Keywords

Research skills, instrument, psychometric properties, dimensionality

Introduction

In the 21st century, enabling students “to use scientific concepts and methods to address problems in research, professional practice, and daily life” (Opitz, Heene, & Fischer, 2017, p. 1) is considered an important goal. In other words, education aims (among other things) at empowering students to acquire a broad set of “skills and abilities to understand how scientific knowledge is generated in different scientific disciplines, to evaluate the validity of science-related claims, to assess the relevance of new scientific concepts, methods, and findings, and to generate new knowledge using these concepts and methods” (Fischer et al., 2014, p. 29). In this paper, this broad set of skills is referred to as research skills. As an indication of this international trend, the emphasis on research skills is reflected in highly valued outcomes in secondary and higher education (Engelmann, Neuhaus, & Fischer, 2016; Fischer et al., 2014; Osborne & Dillon, 2008). Accordingly, in the Flemish educational guidelines for the general track¹, which “focuses on broad general education” (EACEA, 2018, p. 2.2) three specific objectives related to research are formulated for students in the 11th and 12th grade. These curriculum standards state that students have to be able (1) to orientate themselves towards a research problem by collecting, organizing and editing information in a well-considered way; (2) to prepare, carry out and evaluate research, and (3) to report the research results and conclusions and confront them with other points of view (Departement Onderwijs en Vorming, 2017).

In order to explore ways to effectively support the acquisition of research skills, first one needs to find a reliable way to assess these skills (Opitz et al., 2017). In this paper, it is argued that, given the importance of the development of research skills in general secondary education (as reflected by Flemish and international curriculum standards) and the lack of instruments to assess these skills in particular contexts (and, more specific, in a behavioral sciences context), there is a need for a test measuring research skills in a broad (covering all subskills theoretically linked to research skills) but domain-specific way (Fischer et al., 2014). Therefore, drawing upon the theoretical framework of Fischer et al. (2014), a new instrument for this specific target group is developed, meeting the criteria for internal consistency reliability, content and construct validity (Leuven Research Skills Test, further on referred to as LRST). The main focus of this study is to examine the psychometric properties and the underlying dimensionality of the LRST.

In what follows, first, the term research skills is conceptualized. Second, existing tests assessing research skills are identified and evaluated in view of this study. Third, the development process of the LRST, and the format of the resulting instrument is clarified. Fourth, based on confirmatory factor analyses, three models with a different

¹ In Flanders, four different types of education are offered from the second stage of secondary education onwards (EACEA, 2018) (general secondary education, technical secondary education, secondary education in the arts and vocational secondary education).

dimensional structure (based on the review of Opitz et al., 2017) are presented and compared in view of their model fit. To conclude, the results regarding the psychometric properties and the dimensional structure of the LRST are discussed.

Assessing research skills

Conceptualization. Despite the (international) consensus that the acquisition of research skills is an important goal of secondary education, there still exist numerous terms referring to variants of this complex set of skills (Koslowski, 2012; Kestens, Elen, & Verburgh, 2016), such as scientific reasoning skills (Engelmann et al., 2016; Fischer et al., 2014; Opitz et al., 2017; Zimmerman, 2000), scientific literacy (Norris, Phillips, & Burns, 2014) or research methods skills (Earley, 2014). In the present study, the term research skills is used, as it adequately reflects the target concept as a broad set of skills (not merely referring to reasoning, literacy or research methods skills). In addition, many researchers seem to focus on different aspects of research skills (Koslowski, 2012). In order to operationalize this broad set of skills (covering all subskills theoretically linked to research skills, see: instrument development), this study draws upon the theoretical framework of Fischer et al. (Fischer et al., 2014; Opitz et al., 2017), labeling research skills as a comprehensive set of eight scientific activities in which students have to be able to engage, namely: (1) problem identification, (2) questioning, (3) hypothesis generation, (4) construction and redesign of artefacts, (5) evidence generation, (6) evidence evaluation, (7) drawing conclusions and (8) communicating and scrutinizing (Table 1 offers a short description of each of these skills).

Table 1

Conceptualization of research skills (based on Fischer et al., 2014; Opitz et al., 2017)

Skill name	Skill description
Problem identification	During problem identification, a problem representation is built from an analysis of the situation
Questioning	During questioning, an applicable research question is formulated or identified
Hypothesis generation	The process of hypothesis generation involves formulating the hypothesis according to scientific standards (Fischer et al., 2014, p. 34) and a good understanding of the relationship between variables
Construction and redesign of artefacts	This skill includes the construction of some kind of artefact (for example interview guidelines or a questionnaire). This artefact is tested in an authentic environment (Kolodner, 2007)
Evidence generation	Evidence generation includes various approaches such as observation, experimenting, interviewing or surveying
Evidence evaluation	During evidence evaluation, pieces of evidence are evaluated against the available data, or against the methods used to collect the evidence
Drawing conclusions	This skill involves the critical analysis and interpretation of data
Communicating and scrutinizing	This skill involves the use of objective, scientific language to share findings

Existing test instruments. Although to date several tests have been developed to assess students' research skills (Opitz et al., 2017), their conceptual base and the skills they address differ widely (Gormally, Brickman, & Lut, 2012; Osborne & Dillon, 2008). In their review, Opitz et al. (2017) analyzed several tests measuring research skills, with respect to the following features: target group(s), theoretical background, psychometric properties, test format and the skills addressed.

In view of the current study, a first important finding of the review of Opitz et al. (2017) is that only two tests were found to measure research skills in a behavioral sciences context (Frederiksen & Ward, 1978; Heene, 2007). It is observed that most instruments focus on specific natural science domain contexts, mainly in biology, chemistry and physics (Opitz et al., 2017). The authors argued that, "if you only want to measure scientific reasoning in relation to one subject area, it is advisable to choose a test that is set in this specific context and aims to measure scientific reasoning in a domain-specific way (e.g., Cloonan & Hutchinson, 2007; Hammann et al., 2008b)" (Opitz et al., 2017, p. 17). As such, for the specific purpose of this study, there is a need for a test focusing on the behavioral sciences domains. Therefore, the following two comments refer to the two tests developed for a behavioral sciences context

(Test of Scientific Thinking, Frederiksen & Ward, 1978; Empirical-based reasoning, Heene, 2007). Both tests target university students and only use open-ended questions.

As a second important finding, Opitz et al. (2017) argued that an explicit aim of tests should be to include all the skills that are perceived relevant in the scientific reasoning conceptualization that is used (Opitz et al., 2017). It is noticed that the two tests for behavioral sciences (Frederiksen & Ward, 1978; Heene, 2007) only focus on respectively three (hypotheses generation, evidence generation, communicating and scrutinizing) to two (evidence generation, evidence evaluation) skills (Opitz et al., 2017). In view of the specific purpose of this study, it is thus observed that a broad domain-specific test for behavioral sciences, covering all skills mentioned in the conceptualization used in this study (Fischer et al., 2014), is lacking.

Third, it is noticed that the target groups in the two existing tests differ from the target group in this study (students in 11th and 12th grade). The two studies focus respectively on candidates for admission to graduate school in psychology, often senior psychology major students planning to attain a doctorate (Frederiksen & Ward, 1978), and on students with a mean age of 23 (Heene, 2007). As such, the adaptation of these tests to the (much younger) sample of this study might cause problems, such as floor effects. In addition, the specific goals of the tests differ from the goal of this study (developing a test as an instrument for intervention studies). In the case of Heene (2007), the test served as a subtest (with one open question on how to empirically investigate a specific casus) in a selection instrument for higher education. In the case of Frederiksen and Ward (1978), the purpose was to develop a set of tests that could serve as criterion measures in subsequent investigations of creative thinking.

As a fourth comment, Opitz et al. (2017) observed that, next to psychometric properties (such as internal consistency reliability), it should also become a standard procedure to check dimensionality as a construct validity measure. The authors pointed out that, overall, results from factor analyses on current tests are few in number and differ widely in their outcomes (ranging from 1- to 11-factor models). In addition, the number of factors depends on the number of skills included in the test. Overall, Opitz et al. (2017) concluded that a multidimensional structure seems to be slightly more probable compared to a unidimensional one. However, a conclusive answer to the question of dimensionality, and the exact number of research skills factors, is lacking at the moment. Therefore, the authors argued that “as long as we cannot exclude the possibility that scientific reasoning is multidimensional, we should measure several aspects of the construct if we want a broad assessment because we cannot assume that a single sum score will accurately predict the performance of subskills” (Opitz et al., 2017, p. 17). In addition, because of the conceptual differences, the different aims (Frederiksen & Ward, 1978), or the single-item format (Heene, 2007) of the existing tests for behavioral sciences, no clear factorial structures (interpretable in this study) can be deduced. As such, for the specific objective of the present study, neither test (for behavioral sciences, nor in other contexts) is satisfactory. Thus, based on the important comments mentioned above, there is a need for a new instrument with good psychometric properties and clear construct validity, measuring research skills in a broad way, suitable for a behavioral sciences context, further on referred to as the Leuven Research Skills Test or LRST.

Instrument development

The development process of the LRST followed four phases outlined in the Standards for Psychological and Educational testing (AERA [American Educational Research Association], APA [American Psychological Association], & NCME [National Council on Measurement and Education], 2014).

Purpose and scope. In a first and a second phase, the purpose of the test and the scope of the domain to be measured was delineated, and the test specifications were developed and evaluated. The purpose was to construct an instrument (1) assessing research skills (2) designed for students in 11th and 12th grade, (3) with a domain-specific focus on behavioral sciences (4) in view of testing the effects of interventions. Although the current administration takes place in Flemish schools, the idea of this study is that, if successful, the test can be used in schools beyond Belgium. As indicated above, the intended use of the instrument is to assess research skills in an internal consistent way, in order to be able to design instruction based on intervention studies. Establishing content validity was an important part of the development process. Gormally, Brickman and Lutz (2012, p. 365) described content validity as “the extent to which the instrument measures all facets of a given social construct”. As such, the present study refers to the eight skills mentioned by Fischer et al. (2014, see Table 1) as eight essential facets of research skills. Several experts in the field of psychometrics or in the field of complex skills revised whether all eight subskills were represented in the LRST. In addition, the alignment with the three educational objectives related to research skills in 11th and 12th grade of general education (as the described in the introduction) was confirmed by the reviewers.

Item construction. In a third phase, the test items were developed, scoring guides and procedures were formulated, followed by a field testing and evaluation phase. Based on the theoretical framework (Fischer et al., 2014), the eight skills were articulated as outcomes measurable by means of multiple choice or short open answer questions.

The format of the question (multiple choice or short open answer) depended on the characteristics of the skill to be measured (for evidence evaluation, for example, a short argumentation is needed, while this is not necessary when a student has to organize several parts of an abstract). Because expert opinions can promote content validity, the items were reviewed by several experts. First, an existing test previously used within the research group (the LOC-test) (Sermeus et al., 2015) was discussed with the test developers (experts in the field of research skills and critical thinking). Since this test had a low internal consistency score (Sermeus et al., 2015), several problems of the LOC-test (such as bottom effects, misinterpretations or small variances between student scores) were discussed. Second, the same experts reviewed the items of the LRST. These three experts, and three co-authors* of this paper, individually read and commented on the items. Comments were made on the question formats, the difficulty level of the questions, and possible interpretation problems were addressed. Based on this feedback, the test was revised. Third, a cognitive interview was administered with three adolescents, aged 16-18, who studied or had studied in the general track¹. Two students had studied *behavioral sciences*, and one student studied *mathematics-sciences*. Cognitive interviewing can play an important role answering the validity question (Castillo-Díaz & Padilla, 2013). A probing method was used in order to explore the respondents' interpretation processes while answering the questions (Collins, 2003). For example, the students were asked to rephrase the items in their own words (comprehension), and to explain their answers (response) (Collins, 2003). Based on the feedback received during these different phases, modifications were made to various test items.

Field testing. In the fourth phase the test was assembled, piloted and prepared for a large-scale assessment (see: Method) (Adams & Wieman, 2011). In November 2017, the test was administered in a field testing during a class period of 100 minutes. In total, 80 11th grade students, and 88 12th grade students in a behavioral sciences-track (nested in ten classes, in five schools) participated in the study. In addition, several recommendations were made by an expert in test construction and psychometrics. Again, comments were made on the question formats, the difficulty level of the items and possible interpretation problems were addressed. The LRST was revised accordingly. An extensive overview of the items in the LRST can be found in the method section of this paper (Table 2).

The current study. The purpose of the present study is to investigate the psychometric properties and the underlying dimensionality of the LRST. To this end, the LRST was administered to 405 students in 11th and 12th grade. Next, a confirmatory factor analysis was performed to confirm the construct validity of the LRST. Based on the overview of Opitz et al. (2017), three models were compared considering their model fit indices: a unidimensional model with one factor, a multidimensional model with eight factors, and a hierarchical model with eight subordinate factors and one single uniting upper level factor. In addition, the internal consistency reliability was calculated.

Method

Participants. The LRST was administered to 405 students in 11th (41%) and 12th (59%) grade. The cohort consisted of students in four types of secondary education: (a) general secondary education, behavioral sciences track; (b) general secondary education, non-behavioral sciences track; (c) technical secondary education, and (d) secondary education in the arts. Because of the specific interest in students in the behavioral sciences track, a purposive sample was used with 55% students coming from the behavioral sciences track in general secondary education. In order to ensure variation, the other students in the sample followed other tracks, namely other tracks in general secondary education (32%), technical secondary education or secondary education in the arts (13%). The sample consisted of 68% girls, which is in line with the expectations of prevalence of sexes based on the formulated strata (since the behavioral sciences track in Flanders is mainly populated by girls).

Instrument. The LRST contains 59 items, each related to a (sub)skill (see Table 1, Fischer et al., 2014). Each item (in all three item formats) was scored as 0 or 1, relying on a strict scoring guide. Students could not get partial credit (for example, answers in the multiple choice with clarification format were only scored as 1 if the multiple choice answer and the clarification were correct). The total score on the LRST is calculated by adding all eight mean subscale scores, and dividing them by the number of subscales. The number of items per skill, their item format and item examples are visualized in Table 2. The corresponding author scored the test, and a second rater independently coded the answers of 10% of the students to check interrater reliability.

Procedure. Completing the test took maximum 100 minutes. The LRST was administered under supervision, by means of an online web survey. An informed consent form notified each participant about the duration of the test and the specific goal of the study.

Table 2
Test design of the LRST

Skill name	Multiple choice questions	Multiple choice question with clarification of the chosen option	Short open answer questions	Item example
Problem identification			8	Read the two fragments below. a. Write down two common ideas b. Write down one unique idea for each fragment. (Texts about gender and ethnic diversity in relation to academic performance)
Questioning	7			Indicate the best formulated research question for each situation. A researcher is interested in the relationship between grade retention and students' motivation a. How does grade retention influence students' motivation and how do they perceive this? b. Does grade retention have positive effects on students' motivation? c. How motivated are students after grade retention? d. What is the relationship between grade retention and students' motivation?
Hypothesis generation			6	Identify the dependent and independent variables in the situation below. Students are divided in three groups. One group plays math games on an Ipad. One group plays math games on paper, and one group is instructed by a teacher. At the end of the day, the teacher wants to find out which group scores best on a math test.
Construction and redesign of artefacts		4		Twenty persons working in a company are interviewed concerning their use of e-mail. Identify two questions you would change in view of reliability and validity, and adapt them.
Evidence generation	8			Select the best fitted method for data collection for each situation. Karel wants to know more about the opinion of 80 teachers in his school on the use of cellphones at the playground. a. Survey b. Interview c. Experiment d. Observation
Evidence evaluation	4	1	2	Sidra interviews pupils with a gaming addiction. Because she has a lot of respondents, she asks her friend Ruben for help. Sidra interviews half of the respondents, and Ruben interviews the other half. After the interviews they conclude: "we can recognize two types of game addicts. 78 persons talk openly about their addiction, and 69 persons struggle doing this." Give two reasons why this conclusion may be flawed.
Drawing conclusions	1	4	1	Which scatterplot shows the results of the table above? (scatterplot with scores on a test and on an exam, and a table with the same scores is provided)
Communicating and scrutinizing	2	7	4	Some students are conducting research on students' wellbeing in their school. The abstract of their research is noted below, but the fragments are formulated in a random order. Indicate the correct order.

Data analysis. As Gormally et al. (2012) indicated "construct validity involves statistical analyses to evaluate item validity and relationships between instrument items" (p. 366). Therefore, first, an item analysis was done on the 59 items (see Table 3). Item-scale correlations (to check item consistency within each subscale) and *p*-values (to investigate item difficulty) were examined and reported. Second, using the lavaan package (Rosseel, 2012) in R, a series of confirmatory factor analyses (estimation method: WLSMV) was carried out on the 59 items, in order to investigate the construct validity of the LRST. The confirmatory factor analysis method was chosen (instead of an exploratory method) because of the specific interest in the *testing* of the structures of three hypothesized models (Field, 2013). Thus, it is a theory-driven (Schreiber et al., 2006) technique allowing to test and compare several models. Three models were specified based on the overview of Opitz et al. (2017). Model 1 is a unidimensional (single factor) model, with one latent factor (research skills) and 59 observed variables (items). Model 2 is a multidimensional (eight-factor)

model, with eight latent factors (see eight subskills in Table 1) and 59 observed variables (each related to one latent factor, see Table 2). Model 3 is a hierarchical model with eight subordinate latent factors (as described above), one single uniting upper level latent factor (research skills) and 59 observed variables (each related to one latent factor, see Table 2). The adequacy of these three competing models was assessed through an examination of a variety of fit indices. Although the most frequently used fit index is the model's χ^2 (if significant, χ^2 suggests that the data differ significantly from the hypothesized model), it is argued that 'trivial' variations in fit can easily produce a statistically significant χ^2 statistic, especially with large sample sizes (Chen, Curran, Bollen, Kirby & Paxton, 2008; Hu & Bentler, 1999). Therefore, the three models were compared based on other model fit indices and cutoffs (a cutoff value close to .95 for TLI and CFI; a cutoff value close to .08 for SRMR, and a cutoff value close to .06 for RMSEA) identified by Hu and Bentler (1999). However, as will be further pointed out in the discussion section, the use of these cutoff values in personality and ability research are not without controversy, since correlations between items are typically low or moderate in this type of research, often leading to low or moderate factor loadings (Beauducel & Wittmann, 2005). In addition, the critics on the use of cutoff values in any case are pointed out in the discussion section (Chen et al., 2008; Yuan, Chan, Marcoulides, & Bentler, 2016). Third, the internal consistency reliability measures were calculated for each of the subskills of the multidimensional models, and for the total test. Because it is not assumed that the loadings of the items are equal in the factor analysis, (and thus tau equivalence is not assumed), Omega total is used as a measure for internal consistency reliability (McNeish, 2017).

Results

The difficulty level of the items varies between 15% and 89%. The overall test mean is 57% (based on 59 items). Questions related to hypotheses generation and evidence evaluation yield the lowest results, while items related to evidence generation and problem identification yield the highest results. The means, item format and item-scale correlations are outlined in Table 3.

Results on the model fit indices described in the Method section are summarized in Table 4. The single-factor model (Model 1) proved to be the poorest fit to the data, with CFI and TLI values much below the cutoff point of .95 (Hu & Bentler, 1999). Furthermore, RMSEA and SRMR values are too high. The multidimensional eight-factor model (Model 2) proved to be a better fit to the data than Model 1 for all model fit indices. The hierarchical model with eight first-order factors and one overall second-order factor (Model 3) proved to be the best fit to the data. Overall, ranking the models in terms of their overall fit suggests that a hierarchical model (Model 3) performed best.

Modifications to the best fitted model (Model 3) were investigated, as long as content validity was ensured by the remaining items. Items without significant (or with very low: $< .40$) factor loadings were discarded from the model (one item related to problem identification, five items related to questioning, five items related to evidence evaluation, two items related to evidence generation, two items related to drawing conclusions and seven items related to communicating and scrutinizing). The item analysis helped to understand why these items loaded low on the factor (one or several problems with these items, such as low item-scale correlations and/or very low or very high p-values had been noticed during the item analysis). However, content validity was not flawed by deleting these items since the remaining items were representative for the skill. Despite low factor loadings, two items were retained in the definitive model because of theoretical reasons: in view of content validity, the authors chose to retain item EG3 and EG5². As such, all items deemed relevant are included in the subscale.

As Schreiber et al. (2006) indicate, a good fit is probable if the vast majority of the indexes are good. Thus, for the 37 remaining items, overall, the adapted model showed good model fit results ($\chi^2 = 901.259$, $df = 621$, $p < .001$, $CFI = .94$, $TLI = .94$, $RMSEA = .03$ with a 90% confidence interval = $[.03, .04]$, $SRMR = .09$) (Hu & Bentler, 1999). The factor loadings can be found in Table 5.

The internal consistency reliability of the subscales of Model 4 was calculated by means of Ordinal Omega Total, and was .79 for problem identification; .90 for hypotheses generation; .71 for questioning; .71 for evidence generation; .93 for construction and redesign of artefacts; .66 for evidence evaluation, and .70 for communicating and scrutinizing. Due to a limited number of items in the scales questioning and evidence evaluation, Omega total could not be computed for these scales. For the total test, based on 37 items, Ordinal Omega total was .87. The interrater reliability, which was measured with Cohen's Kappa, was $\kappa = .89$ for the total test, and the κ 's for the subskills varied between .70 and .96.

² In the other EG-items, students have to select surveying and interviewing as adequate methods for data collection, while in EG3 and EG5, students need to select experiment and observation as adequate methods for data collection.

Table 3

Item analysis

Skill name	Number of items	Item name	Item format ¹	p-values	Item-scale correlation
Problem identification	8	PI1	CR	.82	.26
		PI2	CR	.60	.40
		PI3	CR	.57	.49
		PI4	CR	.63	.48
		PI5	CR	.80	.24
		PI6	CR	.53	.40
		PI7	CR	.61	.30
		PI8	CR	.62	.31
Hypotheses generation	6	HG1	CR	.33	.51
		HG2	CR	.45	.66
		HG3	CR	.41	.64
		HG4	CR	.51	.61
		HG5	CR	.49	.55
		HG6	CR	.43	.55
Questioning	7	Q1	MC	.60	.22
		Q2	MC	.70	.23
		Q3	MC	.41	.09
		Q4	MC	.58	.26
		Q5	MC	.15	.06
		Q6	MC	.49	.05
		Q7	MC	.67	.07
Evidence generation	8	EG1	MC	.72	.34
		EG2	MC	.29	.04
		EG3	MC	.87	.11
		EG4	MC	.89	.32
		EG5	MC	.75	.21
		EG6	MC	.82	.12
		EG7	MC	.82	.18
		EG8	MC	.83	.32
Construction and redesign of artefacts	4	CA6	MCCR	.45	.52
		CA7	MCCR	.63	.63
		CA8	MCCR	.63	.71
		CA9	MCCR	.73	.66
Evidence evaluation	7	EE1	MC	.64	.06
		EE2	MC	.63	.09
		EE4	MC	.37	.02
		EE5	MC	.17	.08
		EE10	CR	.25	.11
		EE11	CR	.38	.12
		EE12	MCCR	.29	.05
Drawing conclusions	6	DC1	MCCR	.35	.18
		DC2	MCCR	.72	.22
		DC3	CR	.68	.26
		DC4	MC	.69	.33
		DC5	MCCR	.51	.27
		DC6	MCCR	.21	.28
Communicating and scrutinizing	13	CS1	CR	.58	.37
		CS2	CR	.57	.32
		CS3	CR	.47	.41
		CS4	CR	.61	.32
		CS5	MCCR	.82	.16
		CS6	MCCR	.74	.11
		CS7	MCCR	.89	.15
		CS8	MCCR	.60	.03
		CS9	MCCR	.69	.15
		CS10	MCCR	.45	.23
		CS11	MCCR	.58	.30
		CS12	MC	.26	.07
		CS13	MC	.47	.25

¹“CR”: Constructed response; “MC”: multiple choice; “MCCR”: multiple choice with constructed response as argumentation

Table 4

Results of confirmatory factor analysis (CFA) based on 59 items (Model 1, 2 and 3) and 37 items (Model 4)

Model	χ^2	df	p-value	CFI	TLI	RMSEA	SRMR
1	4204.905	1652	.00	.39	.37	.06; [.06-.07]	.16
2	1925.171	1624	.00	.93	.92	.02; [.02-.03]	.09
3	1937.301	1644	.00	.93	.93	.02; [.02-.03]	.09
4	901.259	621	.00	.94	.94	.03; [.03-.04]	.09

Table 5

Item loadings of CFA

Second order factor (Research Skills)								
	PI	HG	Q	EG	CA	EE	DC	CS
	.68	.30	.56	.62	.57	.74	.70	.56
First order factors								
	PI	HG	Q	EG	CA	EE	DC	CS
PI1	.43							
PI2	.62							
PI3	.77							
PI4	.77							
PI6	.65							
PI7	.47							
PI8	.44							
HG1		.71						
HG2		.93						
HG3		.81						
HG4		.90						
HG5		.88						
HG6		.87						
Q1			.59					
Q2			.55					
EG1				.50				
EG3				.38				
EG4				.60				
EG5				.42				
EG7				.56				
EG8				.78				
CA6					.70			
CA7					.86			
CA8					.94			
CA9					.92			
EE10						.50		
EE11						.46		
DC3							.45	
DC4							.68	
DC5							.60	
DC6							.59	
CS1								.93
CS2								.55
CS3								.90
CS4								.50
CS10								.49
CS11								.61

Note. Standardized coefficients are reported. CFA = confirmatory factor analysis; PI = Problem Identification; HG = Hypothesis generation; Q = Questioning; CA = Construction and Redesign of Artefacts; EG = Evidence Generation; EE = Evidence Evaluation; DC = Drawing Conclusions; CS = Communicating and Scrutinizing.

Discussion and conclusion

This study aimed to develop and evaluate an instrument to measure 11th and 12th grade students' research skills in a behavioral sciences context: the Leuven Research Skills Test (LRST). The final LRST consists of 37 items, each

related to a (sub)skill (Fischer et al., 2014, see Table 1). In this article, the psychometric quality and the factorial structure of the LRST, administered to 405 11th and 12th grade students, is investigated.

The item analysis revealed that the LRST can be considered as an achievement test with medium-level difficulty. Additionally, relying on the theoretical framework of Fischer et al. (2014) and drawing upon the Flemish educational guidelines, the content validity of the LRST was established. Furthermore, concerning construct validity, results from factor analyses supported a hierarchical model with eight subordinate latent factors (corresponding to the eight subskills of Fischer et al. (2014)), one single uniting upper level latent factor (research skills) and 37 observed variables (each related to one latent factor), producing the closest fit to the data (compared to a unidimensional and an eight-dimensional model). The inadequacy of the unidimensional model is similar to the results reported by Opitz et al. (2017), arguing a multidimensional model to be more probable. After item deletion (based on low factor loadings), the hierarchical model produced reasonable model fit. Furthermore, results from the internal consistency reliability analysis revealed a good value for the total LRST (Ordinal Omega total = .87). In addition, the internal consistency reliability coefficients varied between .66 and .93 for the subscales.

The results provide initial evidence for using the LRST as a valid measure for research skills proficiency with 11th and 12th grade students in general tracks. In addition, the LRST can be administered in order to further investigate individual differences in research skills proficiency. Furthermore, the LRST can be used as a tool in intervention research, in order to measure the effects of experimental instructional interventions aiming to foster research skills.

Despite these important theoretical and practical implications, some important limitations of this study need to be addressed. First, it should be noted that, because of item deletion, two scales only consist of two items. Future research should try to optimize or construct (new) items to make a stronger scale.

Second, despite the fact that the hierarchical model was found to be the best fit to the data, and all model fit indices reach (or are very close to) the cut-off values formulated by Hu and Bentler (1999), one could argue that the CFI and TLI values are still .01 too low (.94 instead of .95), and that the SRMR-value is .01 too high (.09 instead of .08). However, it is argued that the fit indexes used in this study, initially developed to overcome limitations of the χ^2 statistic (Beauducel & Wittmann, 2005), are based on studies with loadings between .70 and .80 (Hu & Bentler, 1999). As Beauducel and Wittmann (2005) indicate, in psychometric research in many areas of psychology (because of its complexity and its non-exhaustive character), lower loadings (between .40 and .60) are not rare in CFA. Therefore, Raykov (1999) suggests the use of RMSEA in such research, because RMSEA does not assume a perfect (and thus unrealistic) model fit. It is clear that, in this study, the confidence interval of the RMSEA is situated below the cutoff-point of Hu and Bentler (1999). However, it should be noted that the use of cutoff values is criticized in recent literature by Yuan et al., 2016, although these authors do not want to propose an alternative set of cutoff values since these would be sensitive to similar disadvantages. As such, Chen et al. (2008) question whether any fit indices are needed to evaluate a model, but argue that, if used, one must never rely on a single fit index but examine multiple indices instead. As such, the authors of this paper used several fit indices in combination with human judgment (Chen et al., 2008) when reaching the final decision about the evaluation of the model fit.

Third, the question must be raised whether the superiority of the eight-factor and the hierarchical solution (compared to the one-factor solution with one scientific reasoning skill) results purely from the underlying dependence of the items (measuring the same construct), or also results partly from an alternative explanation regarding question format. The fact that the items within a factor are similarly formatted, can contribute to the existence of eight distinct factors. As such, these questions call for further investigation.

Fourth, a hierarchical model (compared to a unidimensional and an eight-factor model) seems to be the best fit to the data. However, because the low number of items for two of the eight subscales, and a rather low internal consistency rate for one subscale, the authors recommend to consider the LRST as an instrument measuring research skills with eight theoretically distinct constructs, but to make no distinction between the subskills at an operational level (yet) (and thus to rely on the total test scores). This recommendation holds until all subscales are optimized, and consist of minimum three items. Furthermore, future research should check for additional construct validity measures. As Boomsma (2000) indicated, testing the validity of the proposed modified models, asks for some sort of cross-validation (ideally with an independent sample from the same population). Although the current sample was too small to cross-validate the model, the authors plan to test the model with an independent sample in the future. Unfortunately, there is no concurrent test to be used for this specific focus in order to establish convergent validity. Still, discriminant validity can be ensured in future studies by means of an IQ-test, or a test measuring language or mathematical skills. To conclude, it would be useful to have some knowledge about the predictive value of research skills proficiency for several measures (such as professional preferences or study success in university) (as also recommended by Opitz et al., 2017).

Acknowledgements

This study was carried out within imec's Smart Education research programme, with support from the Flemish government.

References

- Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education*, 33(9), 1289–1312. <https://doi.org/10.1080/09500693.2010.512369>
- AERA [American Educational Research Association], APA [American Psychological Association], & NCME [National Council on Measurement and Education]. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA Publications.
- Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, 12(1), 41–75.
- Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural equation modeling*, 7(3), 461–483.
- Castillo-Díaz, M., & Padilla, J.-L. (2013). How cognitive interviewing can provide validity evidence of the response processes to scale items. *Social Indicators Research*, 114(3), 963–975. <https://doi.org/10.1007/s11205-012-0184-8>
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological methods & research*, 36(4), 462–494.
- Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, 12(3), 229–238. <https://doi.org/10.1023/A:1023254226592>
- Departement Onderwijs en Vorming. (2017). Secundair onderwijs - Derde graad ASO - Humane wetenschappen - Specifieke eindtermen. Retrieved June 6, 2018, from <http://eindtermen.vlaanderen.be/secundair-onderwijs/specifieke-eindtermen-aso/-/humane-wetenschappen/specifieke-eindtermen.htm>
- EACEA. (2014). Belgium - Flemish Community Overview. Retrieved June 6, 2018, from https://eacea.ec.europa.eu/national-policies/eurydice/content/belgium-flemish-community_en
- Earley, M. A. (2014). A synthesis of the literature on research methods education. *Teaching in Higher Education*, 19(3), 242–253. <https://doi.org/10.1080/13562517.2013.860105>
- Engelmann, K., Neuhaus, B. J., & Fischer, F. (2016). Fostering scientific reasoning in education – meta-analytic evidence from intervention studies. *Educational Research and Evaluation*, 22(5–6), 333–349. <https://doi.org/10.1080/13803611.2016.1240089>
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: SAGE Publications.
- Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dörner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, 4, 28–45. <https://doi.org/http://dx.doi.org/10.14786/flr.v2i2.96>
- Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem-solving. *Applied Psychological Measurement*, 2(1), 1–24. <https://doi.org/10.1177/014662167800200101>
- Gormally, C., Brickman, P., & Lut, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE Life Sciences Education*, 11(4), 364–377. <https://doi.org/10.1187/cbe.12-03-0026>
- Hammann. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education* 42(2), 66–72. <https://doi.org/10.1080/00219266.2008.9656113>
- Heene, M. (2007). *Konstruktion und Evaluation eines Studierendenauswahlverfahrens für Psychologie an der Universität Heidelberg*. (Doctoral dissertation). Universität Heidelberg. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=pdx&AN=0227861&lang=de&site=ehost-live&scope=cite>
- Hu, L. T., & 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>
- Kestens, J., Elen, J., & Verburgh, A. (2016). *Kritisch denken en onderzoekscompetenties: de nexus*. (Unpublished master's thesis). KU Leuven.
- Koslowski, B. (2012). Scientific reasoning: Explanation, confirmation bias, and scientific practice. In G. J. Feist, & M. E. Gorman (Eds.), *Handbook of the psychology of science* (pp. 151–192). New York, NJ: Springer.
- McNeish, D. (2017). Thanks Coefficient Alpha, We'll Take It From Here. *Psychological Methods*. Advance online

- publication. <http://dx.doi.org/10.1037/met0000144>
- National Research Council. (2012). *A framework for K-12 science education: Practices, scrosscutting concepts, and core ideas*. National Academies Press. <https://doi.org/10.17226/13165>
- Norris, S. P., Phillips, L. M., & Burns, D. P. (2014). Conceptions of Scientific Literacy: Identifying and Evaluating Their Programmatic Elements. In *International Handbook of Research in History, Philosophy and Science Teaching* (pp. 1317–1344). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-7654-8_40
- Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning – a review of test instruments. *Educational Research and Evaluation*, 23(3–4), 78–101. <https://doi.org/10.1080/13803611.2017.1338586>
- Organisation for Economic Co-operation and Development [OECD]. (2006). Assessing Scientific, Reading and Mathematical Literacy: A framework for PISA 2006. In *Programme for International Student Assessment* (pp. 3–125). <https://doi.org/10.1787/9789264026407-en>
- Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections*, 13. London: Nuffield Foundation. Retrieved from <http://www.fisica.unina.it/traces/attachments/article/149/Nuffield-Foundation-Osborne-Dillon-Science-Education-in-Europe.pdf%5Cnpapers2://publication/uuid/FA17ED57-71AF-429E-B7E5-D9E33DA4A538>
- Raykov, T., & Marcoulides, G. A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling*, 6, 292–300.
- Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36. URL <http://www.jstatsoft.org/v48/i02/>.
- Schreiber, B., Nora, M., Stage, F., Barlow, E., & King, J. (2006). Reporting structural equatin modeling and confirmatory factor analysis results: a review. *The Journal of Educational Research*, 99(6), 323–338. <https://doi.org/10.3200/JOER.99.6.323-338>
- Sermeus, J., Janssen, E., Claes, E., Depaepe, F., & Elen, J. (2017). Onderzoekskompetenties : een taaie brok voor leerlingen in het secundair onderwijs. *Impuls*, 48(1), 13–20.
- Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., & Lens, W. (2009). Motivational profiles from a self-determination perspective: The quality of motivation matters. *Journal of Educational Psychology*, 101(3), 671–688. <https://doi.org/10.1037/a0015083>
- Verburgh, A., François, S., Elen, J., & Janssen, R. (2013). The Assessment of Critical Thinking Critically Assessed in Higher Education: A Validation Study of the CCTT and the HCTA. *Education Research International*, 2013(1), 1–13. <https://doi.org/10.1155/2013/198920>
- Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3), 319–330.
- Zimmerman, C. (2000). The Development of Scientific Reasoning Skills. *Developmental Review*, 20(1), 99–149. <https://doi.org/10.1006/drev.1999.0497>