Journal of Information Literacy

ISSN 1750-5968

Volume 15 Issue 2

June 2021

Article

Nierenberg, E., Låg, T., Dahl, T. I. 2021. Knowing and doing: The development of information literacy measures to assess knowledge and practice. *Journal of Information Literacy*, 15(2), pp.78–123.

http://dx.doi.org/10.11645/15.2.2795

Knowing and doing: The development of information literacy measures to assess knowledge and practice

Ellen Nierenberg, Research Fellow in Information Literacy, UiT The Arctic University of Norway. Email: ellen.nierenberg@uit.no ORCID: 0000-0001-8666-8092

Torstein Låg, Senior Academic Librarian, UiT The Arctic University of Norway. Email: torstein.lag@uit.no ORCID: 0000-0002-1325-5235

Tove Irene Dahl, Professor of Psychology, UiT The Arctic University of Norway. Email: tove.dahl@uit.no ORCID: 0000-0002-8036-8627

Abstract

This study touches upon three major themes in the field of information literacy (IL): the assessment of IL, the association between IL knowledge and skills, and the dimensionality of the IL construct. Three quantitative measures were developed and tested with several samples of university students to assess knowledge and skills for core facets of IL. These measures are freely available, applicable across disciplines, and easy to administer. Results indicate they are likely to be reliable and support valid interpretations. By measuring both knowledge and practice, the tools indicated low to moderate correlations between what students *know* about IL, and what they actually *do* when evaluating and using sources in authentic, graded assignments. The study is unique in using actual coursework to compare knowing and doing regarding students' evaluation and use of sources. It provides one of the most thorough documentations of the development and testing of IL assessment measures to date. Results also urge us to ask whether the source-focused components of IL – information seeking, source evaluation and source use – can be considered unidimensional constructs or sets of disparate and more loosely related components, and findings support their heterogeneity.

Keywords

higher education; information literacy; knowing and doing; Norway; quantitative assessment

1. Introduction

Information literacy (hereafter IL; roughly, the complex abilities needed to find, evaluate and purposefully use information) is claimed to be important for learning (Oakleaf, 2014), for our empowerment as citizens (CILIP IL group, 2018), for reducing inequality and increasing tolerance (Thompson, 2003, p.1), and for workplace and business success, to mention just a few life domains. The claims that IL is important, particularly in educational contexts, incites a need to measure it, both for educational purposes (e.g. assessments and documentation of student learning), and for research purposes (e.g. to evaluate the effects of educational interventions designed to further it, or to document the specifics of how IL produces the claimed benefits).

At the same time, the concept of IL is itself hard to pin down and seems to be in a state of constant flux. IL scholars and societies have published a number of complex, and at least on the surface, rather different definitions, standards, and frameworks over the last decades (e.g. Association of College and Research Libraries [ACRL], 2000, 2015; Bruce et al., 2006; Bundy, 2004; Coonan & Secker, 2011; SCONUL Working Group on IL, 2011). This introduces challenges for the measurement of IL. If IL is conceived of as a coherent, unitary, and relatively universal construct, with a stable set of "core" features, then it should be possible to develop IL test scales (possibly with subscales) with strong internal consistency, and a structure that reflects the construct's core features. If, on the other hand, IL is a collection of rather disparate and not necessarily correlated components, as asserted by Beile O'Neil (2005, p.51), then this would have two important implications for test development. Firstly, IL tests should then be considered indexes rather than scales, as the latter measure only unidimensional constructs (Streiner, 2003). Secondly, the use of internal consistency measures, such as Cronbach's alpha, would be inappropriate measures of reliability since they rest on an assumption of unidimensionality.

Another question that is seldom addressed in the IL literature is the degree to which what we *know* about IL is reflected in what we *do* when we evaluate and use sources. Although many studies address *either* IL-knowledge *or* IL-practice, few relate them to each other.

The research described in this report therefore serves a three-fold purpose: First, to develop IL-measures for higher education that are applicable across academic disciplines, and that are brief and easy to administer, but still likely to be reliable and to support valid interpretations. Second, to determine whether what students *know* about IL corresponds to what they actually *do* when evaluating and using sources. Thirdly, to help illuminate the question of whether IL, or its source-related facets, should be conceived of as coherent, unitary constructs, or sets of disparate and more loosely related components.

1.1 Definitions and frameworks

IL has had many different definitions since its inception. While specific skills were highlighted in the 1980's (see American Library Association, 1989, par.3), more recent definitions are more abstract and less skill-oriented (see Association of College and Research Libraries, 2015, p.3). The working definition at the foundation of this study combines the concrete and the abstract of previous definitions: 'Information literacy encompasses the knowledge, skills and attitudes needed to be able to discover, evaluate and use information sources effectively and appropriately in order to answer questions, solve problems, create knowledge and learn.' This definition is suitable not only in educational sectors, but also in broader contexts, such as in the workplace, in daily life, and as a responsible citizen. The original, core dimensions of IL, which are as relevant today as in the term's infancy, are easily recognizable in this new definition. Our definition, and this study, emphasises the source-related core aspects of IL – finding, evaluating and using information.

Evidence of the continual evolution of the IL construct is also reflected in various IL standards and frameworks, such as those from the ACRL. Changes between the ACRL's (2000) *IL Competency Standards for Higher Education* [Standards] and its 2015 *Framework for IL for Higher Education* [Framework] reflect rapid developments in both the HE-sector and the information ecosystems in which we otherwise live and work. The amount, complexity and form of information has increased in the last 15 years, making it more difficult to navigate, evaluate and use ethically. In addition, teachers and librarians have greater responsibilities in teaching core concepts of IL and integrating them into curricula, and students have become information creators to a greater extent than previously (ACRL, 2015). Executive director Mary Ellen Davis further explains:

> One of the primary reasons [for shifting to the ACRL Framework] was that the Standard was seen to reduce learning to a checklist which did not fit well with higher learning concepts that faculty were teaching. The Framework provides a bridge to the faculty for more integrated teaching of the concepts. (personal communication, Feb. 20, 2019)

In addition to the ACRL Standards (2000), other IL standards/frameworks/curricula from before 2015 (including Bundy, 2004; Coonan & Secker, 2011; SCONUL Working Group on IL, 2011), also rely on a checklist of learning outcomes and/or skills. Although less popular after 2015, especially in the US, one advantage of the checklist approach was that it facilitated the assessment of IL competencies, in comparison to the ACRL Framework from 2015, whose abstract «threshold concepts» are harder to measure than concrete learning outcomes (Gross, et al., 2018). A notable feature of all these guides is the broad range of knowledge and skills that are incorporated into the IL concepts.

1.2 Existing IL knowledge tests

There are a number of different ways to assess students' IL. Library instruction and IL teaching is commonly evaluated by asking students to self-report on their learning experiences and/or subjectively estimate their IL (Oakleaf, 2008; Schilling & Applegate, 2012). Such measures, while providing important information, do not adequately capture IL knowledge, skills and behaviour, in part because students tend to overestimate their IL (Gross & Latham, 2012; Ivanitskaya et al., 2006; Julien & Hoffman, 2008; Nierenberg & Fjeldbu, 2015; Oakleaf et al., 2011; Polkinghorne & Wilton, 2010).

Measures of knowledge, usually multiple-choice tests, are another common way to measure IL (see Hollis, 2018; Mahmood, 2017; Walsh, 2009 for reviews). These are less reliant on students' more or less accurate self-assessments, and they tend to be relatively easy to administer and score. Despite their popularity, relatively few of these measures have been psychometrically evaluated (Mahmood, 2017), and some of the more thoroughly evaluated measures are only commercially available and, therefore, less accessible to most IL practitioners.

Evaluations of freely available, general IL knowledge tests typically present validity evidence in the form of documentation of test construction processes that are intended to ensure correspondence between test items and the IL standards or frameworks upon which test development was based (see Beile, 2005; Hollis et al., 2019; Podgornik, et al., 2015 for examples of this approach). To the extent that such frameworks represent consensus on the content and structure of the IL concept, this type of evidence is rightly emphasised.

In designing psychometric tests, factor analyses are often used to explore or identify any latent, underlying variables, called *factors*, from large numbers of observed variables. These factors can then be used to reduce the number of items to those reflecting the major constructs being measured. Interestingly, though, factor analytic investigations of response patterns to IL knowledge tests are rare. We have been able to locate only five. Among these, two find no clear factor structure (Beile O'Neil, 2005; Morley, 2014) and most are done on tests developed to either capture only one circumscribed sub-domain of IL (Catalano, 2015 - source evaluation; Ondrusek et al., 2005 - accessing information in a library context), or to target specific student populations (Beile O'Neil, 2005 – teacher students; Leichner et al., 2014 – psychology students; Morley, 2014 – resident physicians). Furthermore, many of these studies have methodological flaws (e.g. low item to sample ratio, no reports of assumptions checked or rationales for extraction and rotation methods) that weaken them as bases for inference. Lacking any empirical evidence of IL dimensionality is a notable gap. Of course, even though empirical analyses should not be allowed to *determine* our construct definitions, factor analyses can inform the decisions we make regarding how a given construct can or should be conceived (Mulaik, 2009; B. Thompson & Daniel, 1996). Specifically, if one assumes that IL (or the various facets of IL), is a latent variable that causes certain measurable behaviours (see Cronbach & Meehl, 1955), then factor analyses can be used to design a test structured around highly correlated items. If, on the other hand, IL is assumed to be a construct that merely represents a collection of various, possibly quite unrelated indicator behaviours, then factor analyses of test items that represent the concept should not be expected to yield a clear and stable underlying structure.

Another peculiarity of the literature on the development and evaluation of IL tests is the reliance on internal consistency measures of reliability, in particular Cronbach's alpha. Of the 16 studies reviewed by Mahmood (2017), 14 report alphas, while only four report temporal consistency (i.e., test-retest reliability). Given the aforementioned lack of factor analytic evidence regarding the dimensionality of IL tests, and the seemingly pervasive construal of IL as a multifaceted construct in extant definitions and frameworks, this is potentially problematic. If factor analytic evidence, currently lacking for most IL tests, should support unidimensionality (or, alternatively, a clear subscale structure), internal consistency measures make sense. If not, they are inappropriate. In fact, if an IL test is intended to measure a construct that consists of a set of rather different aspects, a high alpha may indicate that the test is too narrowly focused (Streiner, 2003, p.220). Such measures are best considered indexes (whose items measure behaviours that reliably cause the score), rather than scales (which assume an underlying property that reliably causes the behaviours measured by the items and that accounts for their interrelatedness).

Since the assumption of internal consistency may not apply to tests of IL, a more appropriate estimate of reliability may be their stability over time. Such temporal consistency is measured by comparing test results from the same person at different times (test-retest reliability). A reliable instrument should produce similar results if a) there is no intervention between tests, and b) the time interval is suitable. According to Streiner and Kottner (2014, p.1974), 'there should be sufficient time between the two [tests] so that the respondents do not remember their original responses and simply recall those, but not so much time that the construct being assessed can change.' Test-retest reliability is often measured with intraclass correlation coefficient (ICC, Streiner & Kottner, 2014), especially when unidimensionality is not assumed (Thorsen & Bjorner, 2009, p.31). ICC levels are considered adequate if the lower bound of the 95% confidence interval is greater than 0.6 (Multon, 2012).

1.3 The relationship between knowing and doing

Although several IL studies compare students' self-assessments with their competencies (Gross & Latham, 2007, 2013; Smith et al., 2013), there is little IL research attempting to establish a direct connection between their objectively measured IL-knowledge (knowing) and applied skills (doing) in an authentic educational context. Two exceptions are work by Schilling and Applegate (2007) and Beile O'Neil (2005), which both focus on information seeking. Neither of these, however, measure applied skills in actual, mandatory assignments. Although Schilling and Applegate (2007) found that students' IL-knowledge (as measured by an objective test) was *not* related to their subsequent performance on a practical literature search task, Beile O'Neil (2005) found in a similar study that there *was* a significant correlation between the two, although her sample size was small.

1.4 The present study

Three quantitative measures of IL were employed in order to answer the research questions in the present study, and this article describes both the development and testing of these measures, and the implications of the results. The first is a *know*-measure, in the form of a test, and the other two are *do*-measures, based on mandatory student assignments:

1. Know: a multiple-choice IL knowledge test to capture knowledge of three key, source-focused aspects of IL (seeking information, evaluating and using sources)
2. Do: a source evaluation measure to capture students' abilities to select and critically evaluate sources
3. Do: a source use measure to capture students' abilities to use sources correctly when writing

While the multiple-choice knowledge test (hereafter "IL knowledge test") measures what participants *know* about the source-focused components of IL, an annotated bibliography and rubric measure what they *do* in practice when evaluating and using sources, respectively.

The present study differs from previous know-do research in two important ways. Firstly, while Schilling and Applegate (2007) and Beile O'Neil (2005) focused on information seeking, the present study examines associations between knowing and doing in terms of the evaluation and use of sources. Secondly, the present study uses authentic results from mandatory coursework instead of non-graded tasks designed purely for research purposes. Establishing a connection between what the students *know* in theory and what they *do* in practice can be used to test the validity of the IL knowledge test as an indicator of practical abilities related to evaluating and using sources.

Since these tools were designed to measure IL knowledge and skills in higher education (HE), students at different levels were recruited to test their efficacy. An additional survey, measuring respondents' interest in being/becoming information literate individuals (what they *feel*), will be addressed in another article.

In order to explore the possible structure of the IL knowledge test, exploratory factor analyses were undertaken in this study. Given that items were developed to capture three core, source-focused aspects of IL, one might expect extracted factors to align with these aspects. On the other hand, if IL is conceived of as 'one underlying variable' or 'singular construct' (Hollis, 2018, pp.76–77), a latent attribute that causes the observable behaviours that IL tests measure, then one would expect a strong first factor, with remaining factors accounting for very little variance (Streiner, 2003). A third possibility is that a broadly conceived IL test measures a number of rather different facets that may or may not be related to each other in any other way than that they tend to be seen as part of a family of related but distinct IL skills, in which case we should not expect any clear or meaningful factor structure. Given that all of these are reasonable a priori possibilities, exploratory factor analysis seemed the most appropriate analytic approach.

The present study aims to create tools, suitable for multiple disciplines in HE, for measuring IL's source-components, and to psychometrically evaluate these with large sample sizes. The study will also compare what students *know* with what they *do* when evaluating and using sources for real course work. It will also examine the dimensionality of IL, with possible implications for how it should be measured.

2. Methods

2.1 Participants

Data was collected from four different samples, serving slightly different but overlapping purposes. The study was approved by the Norwegian Data Protection Authority, and informed consent was required for participation.

First, for the item selection and wording refinement process, four selected IL experts were recruited to evaluate the validity of the 50 original items. These included a senior lecturer at The Swedish School of Library and Information Science (LIS), and three IL-practitioners in Norway, including the head of the library's Department of Public Services at UiT The Arctic University of Norway, and an assistant professor of education and LIS, and an academic librarian at Oslo Metropolitan University.

Five students participated in think-aloud protocols while answering the test questions. Three were first-year bachelor students and two were college-bound high school seniors (four female and one male). The last two are (or will soon be) new college students – the least experienced of the survey's main target group (undergraduates) – and therefore more likely than higher-level students to have questions/misunderstandings regarding items.

A pilot sample (Table 1) of undergraduates were then recruited in spring 2019 through their Learning Management Systems (LMS). Of these respondents, 40% were in the first year of their current study program. Two universities, with academic studies such as psychology and economics, and one university of applied sciences (UAS), with vocational studies such as nursing and teacher education, were included. All are medium-sized (14,000-18,000 students) public institutions with free tuition. This sample answered all of the 50 initially created items for the IL knowledge test. The data were intended primarily for item selection and test construction purposes.

Table 1: Descriptive statistics: pilot sample – undergraduate (survey in Norwegian)

N	**Semesters completed**	
	M	***SD***
268	3.5	2.58

The first students to complete the final version of the 21-item IL knowledge test were undergraduates at the beginning of their first semester, *before* IL instruction, in 2019 (Table 2). These students had previously completed an average of 0.8 semesters of HE in other disciplines. Respondents were from three universities and one UAS, from south-eastern to northern Norway, in academic and vocational fields ranging from sociology to fisheries management. The institutions are public and small- to medium-sized (5,200-18,000 students). Participants were recruited either in the classroom, via their LMS, or by teaching-librarians from their university libraries, and had the possibility of winning a gift card. The data were intended for assessment of reliability and validity, and for exploring any potential factor structure in the IL test. Many participants (37%) were in their first year of a psychology program at the largest university in our sample. These psychology students provided data for both the IL knowledge test and the two do-measures, by utilizing assignments in their academic skills course.

The undergraduate sample answered the IL knowledge test once again at the end of the semester, *after* IL instruction. They were recruited via e-mail and offered a possible gift card for their participation. During the semester, all received instruction on information searching, and the evaluation and use of sources.

Table 2: Descriptive statistics: undergraduate sample – beginning and end of first semester (survey in Norwegian)

Group	***n***	**Gender**			**Age (*M*)**	**Semesters in HE**	
		Male	**Female**	**Other**		***M***	***SD***
Beginning of first semester							
Psychology students	97	31	66	0	21.3	0.5	0.89
Other students	163	65	98	2	23.4	1.0	1.15
Total	260	95	163	2	22.8	0.8	1.09
End of first semester							
Psychology students	45	14	31	0	20.6	0.6	1.47
Other students	67	16	50	1	23.1	1.5	2.06
Total	112	30	81	1	22.1	1.1	1.89

A graduate student sample of master's and PhD students in a variety of disciplines, from several Norwegian and five international universities, were recruited via social media and email from faculty at other institutions (Table 3). This sample answered the IL knowledge test only once and provided a comparison group for the undergraduate sample.

Table 3: Descriptive statistics: graduate student sample

	n	Gender			Age (*M*)	Language	
		Male	Female	Other		Norwegian	English
Master's	196	57	131	5	32.3	80	116
PhD	170	69	96	2	33.4	65	105
Total	366	126	227	7	32.7	145	221

2.2 Materials

2.2.1 IL knowledge test

2.2.1.1 Item generation

Before generating items for the IL knowledge test, ten international IL frameworks and standards were analysed in order to find one that: a) applies to most disciplines in HE, and b) specifies learning outcomes relevant to central constructs of IL identified in the framework analysis, thereby facilitating assessment. The framework that best fulfilled these criteria was the Australian and New Zealand IL Framework [ANZIL framework] (Bundy, 2004), based on the ACRL's now discontinued Standards (2000). Despite its age, the decision was therefore made to base IL knowledge test items on the ANZIL framework's six standards, regarding respectively: recognizing the need for information, and finding, evaluating, managing, creating, and using information (see Bundy, 2004).

A pool of 50 items (Nierenberg et al., 2021, Pilot) was created with the goal of measuring students' knowledge of aspects of IL described in the framework. Items had varying levels of difficulty and were intended to detect learning over time. All items were multiple choice with four alternative, plausible answers, one of which was correct. 'I do not know' was not an option. An example of a survey item is: 'What is the most important reason to use sources when writing a paper? a) To support arguments; b) To avoid plagiarism; c) To show that you've read the sources; d) To satisfy the requirements of the assignment.'

As discussed below, items were selected from the original pool of 50 items based on expert evaluations of validity and responses to the items in a pilot sample.

2.2.1.2 Expert evaluations of validity

The panel of experts looked specifically at three qualities (headings from Beile, 2005):

1. Clarity: Are items and answer alternatives clear, unambiguous, and understandable?
2. Content accuracy: How well does each item align with the standards and learning outcomes in the ANZIL framework?
3. Objectivity and bias: Are items objective? Are there cultural, geographic, disciplinary, or other biases? (The survey is designed to be used in various contexts and suitable for *all* students, independent of their field of study, location, race, gender, religion, sexual orientation, etc.)

Items were rated by the experts on a scale of 0 (absence of quality) to 3 (fully expressed quality), and those items receiving a score of 2.6 or below were later revised in accordance with experts' feedback.

2.2.1.3 Pilot test

Responses from the pilot sample (*n* = 268) to the 50 original items were analysed to select those most useful for the final survey. In psychometrically designed tests, items should: a) have a suitable range of difficulty, and b) correlate positively to the total test score, i.e., item discrimination (DeMars, 2017). Firstly, items were discarded if more than 88% of the sample answered correctly. Secondly, items with an item-to-total correlation (the point-biserial correlation between the item score and the summed test score) lower than 0.3 were removed. After deleting 38 items based on these criteria, 22 items remained.

Interestingly, these 22 items were all related to information seeking and to the evaluation and use of information sources, thus corresponding to Standards Two, Three and Six in the ANZIL framework (see Bundy, 2004). Items from the other three standards (recognizing the need for information, managing information, and creating new information) were not useful based on the criteria for item difficulty and discrimination. The 22 remaining items were then grouped in terms of difficulty, measured by the percentage of participants who answered the item correctly (P-value): a) easy (P-value 80% - 88%), b) medium (P-value 50% - 79%), and c) hard (P-value < 50%). In cases where multiple items at the same level of difficulty covered the same IL topic, for example several easy items about source evaluation, items with the lowest item discrimination were deleted. When this process was complete there were seventeen items remaining. These items thus demonstrate two characteristics: a) a range of difficulty (without being too easy, to avoid ceiling effects), and b) at least a moderate correlation with total test scores. These seventeen items should therefore be useful in distinguishing between respondents with varying levels of IL.

Four additional items were added to the test at this stage in order to have an equal number of items in the three main categories of information seeking, and the evaluation and use of sources. The final IL knowledge test now consisted of 21 items, four of which had neither been tested in the pilot nor rated by experts.

Since we wanted to make a tool that could be used in multiple languages, the final IL knowledge test was then translated into English, and we tested both language versions with the master's and PhD sample. The original Norwegian version and the English translation are provided in Nierenberg et al. (2021).

Exploratory factor analyses based on data from the pilot sample were also performed at this stage, with the intention to further refine the item selection (see Appendix A) and explore dimensionality of the IL test.

2.2.1.4 Think aloud protocol

In order to test the remaining survey items for readability and comprehension, think-aloud protocols were collected from five current or prospective students. In a think-aloud protocol, members of the target population verbalise their thought processes while performing an action, in this case while taking a survey, saying aloud how they interpret the question and choose their answer (Hughes, 2018, p.767). This can uncover possible misunderstandings, confusion, or ambiguities that were not previously detected. Although think-aloud protocols produce useful evidence for establishing psychometric accuracy, they are nevertheless seldom utilized (Hughes, 2018, pp.764–765).

While taking the IL knowledge test, participants were encouraged to describe verbally how they interpreted the questions, chose their answers, and eliminated answer alternatives. They could also comment on the test in general afterwards. Participants used 15 to 25 minutes to take the survey while thinking aloud. The wording in four items was revised as a result of their feedback.

2.2.2 Annotated bibliography for source evaluation

One of the tools employed to measure what students actually *do*, (as opposed to the IL knowledge test which measures what they *know*), is the annotated bibliography for source evaluation. Several items in the IL knowledge test address the critical evaluation of sources, but in order to see if students actually *use* criteria for source evaluation in practice, an authentic, assignment-based method was desired.

When reviewing the literature, little research was found on measuring source evaluation with quantitative methods; one exception was work by Walton and Hepworth (2012). Their method allows qualitative data from students' annotated bibliographies, in which they evaluate their sources of information, to be assigned quantitative values (Walton & Hepworth, 2012, p.58). A variation of this method was used in the present study. For one of their mandatory assignments, the psychology students from the undergraduate sample chose three academic sources for their term papers and composed references to these sources in a bibliography. Their choice of sources was based on what they had learned about source evaluation during the semester. For each reference, students wrote an annotation describing why it was a good, academic source, and well suited to answering their research question. For the purposes of the present study, annotations from those students who had previously completed the IL knowledge test at the start of the semester (*n* = 93) were analyzed and rated for the following three components:

1. *Quality* of the source: ratings of each source on a scale of 0 (not academic) to 3 (academic).[1]
2. *Variety* of criteria: the number of different, unique criteria for source evaluation stated by the student (e.g. relevancy, accuracy, authority, etc.) in each annotation. Each specific criterion is allocated a score of '1'.
3. *Frequency* of criteria: the number of instances criteria for source evaluation appear in each annotation. Each criterion is allocated a score of '1'. (This will be different from the *variety* score if a criterion, e.g. authority, is listed more than once in the annotation.)

Details about the annotated bibliography measure for source evaluation, with an example annotation, are provided in Nierenberg et al. (2021). All bibliographies were scored by two raters – one of this article's authors and a master's student in psychology. To assure the reliability of their scoring, the raters compared, discussed, and agreed upon assessments for the first ten students' annotations and independently scored the rest. During this process, raters created a list of eligible, "countable" criteria for source evaluation in order to make scoring as objective as possible. Inter-rater reliability for the raters was assessed using ICC when scoring was complete. ICC is the best measure of rater consistency for interval data with a normal distribution, and values over 0.60 are considered acceptable (Bujang & Baharum, 2017; Multon, 2012).

2.2.3 Rubric for source use

After completing the IL knowledge test and the source evaluation assignment, the remaining psychology students (*n* = 87) wrote a term paper on a psychology topic of their choice. In this mandatory assignment, students were required to use a minimum of five academic sources and to cite their sources in APA-style. The assignment was graded using a rubric with 17 evaluation

[1] This is based on the following definition of an academic source used in the introductory psychology course: a) An academic source is written by professionals, for professionals. b) It builds on other academic sources, cites sources to substantiate claims and arguments, and provides complete references for these sources. c) The text has been the subject of a thorough, professional editorial process and has received constructive criticism before being published.

criteria, five of which pertained to the use of sources (see Table 4, and Nierenberg et al., 2021) for more details).

Table 4: Criteria and scoring in rubric for assessing students' use of sources

Criteria for use of sources	No	Partially	Yes
Are academic sources used to support arguments?	0	0.5	1
Are sources cited in the text when necessary?	0	0.5	1
Are the in-text citations written in correct APA-style?	0	0.5	1
Is the reference list written in correct APA-style?	0	0.5	1
Are all in-text citations listed in the reference list, and vice versa?	0	-	1

These five criteria are designed to evaluate the students' abilities to properly use and cite information sources when writing. Three raters, two of this article's authors and a master's student in psychology, evaluated the term papers of the 87 students who had previously answered the IL knowledge test, using this 5-point source use rubric. Raters calibrated their grading beforehand to ensure reliability, and then independently scored the rest. Though APA was our reference style of choice, any reference style can be substituted in its place.

2.3 Reliability and validity evidence

Analyses were conducted in IBM SPSS Statistics 26 (2019) and RStudio (2019).

2.3.1 Reliability analyses

Temporal consistency, measured with ICC test-retest reliability, was calculated from data from 46 undergraduates who took the IL knowledge test twice at the end of the semester. Time intervals varied between 7 and 73 days, with no IL instruction in between, an acceptable test-retest range (Streiner & Kottner, 2014). Inter-rater reliability was assessed for consistency using ICC for both of the do-measures.

2.3.2 Validity analyses

In addition to the comprehensive item selection and refinement process based on pilot testing, expert review and student think-aloud protocols, further support for the validity of the final IL knowledge test would be based on results showing that the overall knowledge test score discriminates among students at different levels of HE, and between undergraduates at the start and end of one of their first semesters. To detect these differences, two analyses were performed. Since the assumption of homogeneity of variance was not met for this data, we used a Welch's ANOVA with Games-Howell post hoc to compare mean IL knowledge test scores for undergraduates (semester-start), master's and PhD-students. In addition, a paired samples t-test on test scores at the start and end of the semester for undergraduate students was performed.

Bivariate correlations were then calculated to find potential relationships between scores on the IL knowledge test and scores on do-measures. Such correlations would provide evidence that what students *know* is related to what they *do* in practice.

2.3.3 Exploratory factor analyses

Given that we constructed items to reflect what might be considered core dimensions of IL, we wanted to explore a possible structure of the IL knowledge measure. We did this not only to aid in item selection, but also to explore whether hypothesised or postulated dimensions of IL were revealed in the structure of the correlations among items. A number of factor models were explored, based on matrices of tetrachoric correlations between items. Factors were extracted using principal axis factoring, and solutions were rotated using ProMax. Details of these analyses are described in Appendix A.

3. Results

All data used in this article is available from UiT Open Research Data (Nierenberg et al., 2021).

3.1 Know: IL knowledge test

Mean scores, standard deviations, and minimum and maximum scores for the IL knowledge test for the different levels of HE are shown in Table 5. For score distribution histograms, see Appendix B.

Table 5: IL knowledge test score: statistics for student groups

Student group	*N*	*M*	*SD*	*Min.*	*Max.*
Undergrad. (semester start)	260	12.46	2.95	3	19
Undergrad. (semester end)	112	13.92	2.97	6	20
Master's	196	16.06	2.49	8	21
PhD	170	16.82	2.11	10	21

3.1.1 Exploratory factor analyses

Factor analyses with pilot data provided no useful basis for decisions regarding item selection. Correlation matrices showed generally poor factorability in all samples, and none of the many potential solutions were readily interpretable. Specifically, neither a one-factor solution, which one might expect if items of the IL knowledge test represented a unitary underlying construct, nor solutions where factors aligned with core facets of IL, or indeed any other meaningful pattern, emerged from these analyses (see Appendix A)

3.1.2 Reliability evidence

Reliability statistics for the IL measures, based on the undergraduate sample, are shown in Table 6. Intraclass correlation coefficients for consistency – for both interrater reliability and test-retest reliability – are all above .80, normally considered sufficient evidence of reliability. Results from graduate students show no difference in mean scores between the Norwegian (*M* = 16.59, *n* = 145, *SD* = 2.33) and English (*M* = 16.29, *n* = 221, *SD* = 2.35) versions of the IL knowledge test (t_{364} = -1.212, *p* = .616). These data are therefore pooled in all subsequent analyses.

Table 6: Reliability statistics for IL measures

IL measure	Reliability measure	Unit	Value	*N*	95% CI	
					LL	UL
IL knowledge test (21 items)	test-retest reliability	ICC	.84	46	.72	.91
Source evaluation measure	interrater reliability[a]	ICC	.89	93	.83	.93
Source use measure	interrater reliability[b]	ICC	.92	87	.89	.95

Note. ICC = intraclass correlation coefficient; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

[a] 2 raters

[b] 3 raters

3.1.3 Validity evidence

The previously described processes of item generation and item selection provide validity evidence for the final IL knowledge test items. In addition, a paired sample t-test for undergraduates' test scores at the beginning and end of their first semester showed a statistically significant difference ($t_{111} = 5.078$, $p < .001$). On average, semester-end scores were 1.21 points higher than semester-start scores (95% CI [0.74, 1.69]) for those students who took both tests.

A Welch's ANOVA showed statistically significant differences between the means of undergraduates (semester-start), master's and PhD scores, $F(2, 623) = 179.49$, $p < .001$. Follow-up procedures, using the Games-Howell post hoc, indicated that the mean score for the undergraduate group ($M = 12.46$, $SD = 2.95$) was significantly different ($p < .001$) from the master's group ($M = 16.06$, $SD = 2.49$), and the mean for the master's was significantly different ($p = .005$) from the PhD group ($M = 16.82$, $SD = 2.11$). Taken together, these results suggest that the IL knowledge test discriminates between students at different levels of HE – more advanced students have higher mean scores. Mean scores with 95% confidence intervals are plotted in Figure 1.

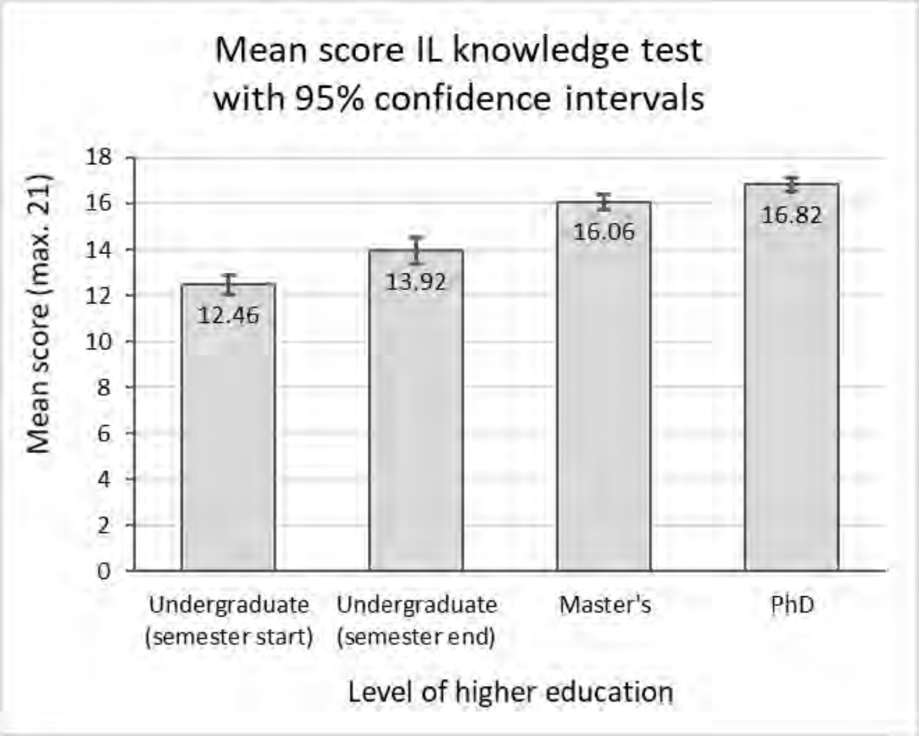

Figure 1: Mean score IL knowledge test with 95% confidence intervals for different levels of HE

3.2 Reliability and validity of assignment-based measures

The interrater reliability for the scoring of both *do*-measures was high (see Table 6). Both measures have inherent validity since teachers devised them to measure aspects of students' IL.

3.3 Know and do: relationship between the IL-test and do-measures

IL knowledge semester-start test scores from psychology students in the undergraduate sample were used in correlation and regression analyses. The correlation matrix and details of regression analyses are found in Appendix C.

3.3.1 Source evaluation measure: annotated bibliography

No significant correlation was found between annotated bibliography total scores and IL knowledge test scores. However, when correlating annotated bibliography component scores for quality, variety and frequency with test scores, a statistically significant, positive, but weak/moderate correlation was found between the quality component score and test score, $r(93) = .27$, $p = .008$. Test scores were not significantly correlated with variety or frequency component scores.

In an attempt to control for HE experience, multiple regressions were performed, using test scores and number of semesters of higher education as predictors, and the quality component score as dependent variable. Unsurprisingly, given the restricted range of the number of semesters variable and its lack of association to the quality score as measured by the simple correlation, it does not influence the predictive strength of the knowledge test scores ($\beta = .27$, $t(91) = 2.70$, $p = .008$). (See Appendix C for details.)

3.3.2 Source use measure: rubric

Bivariate correlation analyses with source use rubric scores indicate a statistically significant, positive, and moderate/weak correlation with IL knowledge test scores, $r(87) = .31$, $p = .004$. Again, we did multiple regressions, using test scores and number of semesters of higher education as predictors, this time using source use rubric score as the dependent variable. This yields a $\beta = .32$, $t(85) = 3.06$ for the test scores when number of semesters is accounted for (see Appendix C).

4. Discussion

The goals of this study were to 1) develop quantitative IL assessment measures for HE that support valid interpretations and have a high degree of reliability, 2) provide these in two languages (Norwegian and English), 3) determine whether there is a connection between what students *know* about IL and what they *do* in practice when evaluating and using sources, and 4) shed light on the question of whether the source-related components of IL can be considered a construct with clear dimensional structure.

To answer these research questions, three measures were developed and examined: a test to measure IL knowledge and two assignment-based tools to measure what students do in practice with their IL knowledge when evaluating and using sources in mandatory coursework. In contrast to self-report IL-measures, which are commonly used to evaluate the impact of IL instruction, the instruments in the present study provide more objective measures of students' IL knowledge and skills.

4.1 Assessment measures

4.1.1 Know: IL knowledge test

The IL knowledge test presented in the current study can provide a valuable contribution to IL-assessment and has several advantages over many other IL tests (see Mahmood, 2017), as demonstrated by the following:

- A thorough item selection process based on framework analysis, pilot testing, expert evaluations and think-aloud protocols, ensures that the IL knowledge test covers three core aspects of IL that are common to nearly all of IL's many definitions/interpretations.
- It has been evaluated for reliability and validity in a number of different samples.
- It is freely available, and quickly and easily administered.
- It is suitable for students in diverse disciplines.

- Reliability was estimated using temporal consistency (test-retest reliability), as opposed to by calculating internal consistency (as most IL test developers do, despite its relevancy only for unidimensional constructs).
- Scores on the IL knowledge test are nearly identical in the English and Norwegian versions, suggesting that it can be used in both languages.
- It discriminates reasonably well between undergraduate students at the beginning and end of their first semester, and between students at different levels of HE.

That only items regarding the source-related components of IL were found useful in measuring IL knowledge was an unexpected result of the item selection process. This conformed with our intent to base the IL knowledge test on those aspects of IL common to its various definitions, and underscores that finding, evaluating and using information are fundamental to the concept of IL.

4.1.2 Do: assignment-based measures

The source evaluation and source use measures utilized in the present study can prove valuable both in quantitatively assessing two core aspects of IL – the evaluation and use of sources – and in examining the connection between knowing and doing in an educational context. Both measures can be scored with high interrater reliabilities. Arguably, they have inherent validity, since they were devised by faculty to assess aspects of students' IL explicitly valued in a real HE context. Importantly, they directly assess performance on practical IL tasks, rather than «mere» knowledge. While there is some evidence that knowing and doing are related (see section 4.2), the two are clearly not equivalent.

Results indicate that only the quality component of the source evaluation measure (ultimately the most important) has significant correlations with IL knowledge test scores at the start of students' first semester. The relevance of source evaluation variety and frequency is worthy of additional scrutiny with other student populations or at different points of the HE experience. Still unknown is if student low variety and frequency scores are a result of an inability to discern how they arrive at their quality appraisal, or their inability to articulate this.

4.2 Knowing and doing

Our study complements the paucity of IL literature regarding the connection between IL knowledge and skills in areas other than information seeking. The present study differs from similar research, for example Beile O'Neil (2005) and Schilling and Applegate (2007), by including two core aspects of IL – the evaluation and use of sources – and by assessing students' skills using authentic, graded coursework. Results provide varied evidence of the connection between what students know and actually do. Scores on the IL knowledge test had positive, significant, and borderline weak-moderate correlations with source evaluation and source use scores. This may be due to the students' short experience with HE and indicate that those who are early in their academic careers are not yet able to effectively use their knowledge in practice. A future study will measure these correlations toward the end of the students' undergraduate program to determine how the ability to translate knowledge into practice evolves with HE experience.

The modest strength of these associations indicates that students' demonstrated practice is strongly influenced by factors other than their knowledge of IL. Assessments of students' IL should therefore be based on multiple measures covering both knowledge and practical performance, and not exclusively on IL knowledge tests. We attempted to control these associations for amount of HE experience, however the range of this variable was rather restricted in our sample, so the control was weak. While amount of HE nevertheless correlated with IL test scores, it did not correlate with any of the do-measures in our first-semester student sample.

4.3 Dimensionality

There is little consensus among IL researchers about the dimensionality of the IL construct, although its homogeneity is often implied. Few studies, with the exception of Beile O'Neil (2005) and Morley (2014), have investigated the dimensionality of IL as a whole using factor analyses, and both concluded that IL is multidimensional. Hollis (2018), on the other hand, explicitly postulates that IL is a unidimensional construct, but has yet to provide factor analytic evidence of this.

Rather than investigate IL as a whole in this study, we have focused on its source-related components – finding, evaluating and using information. Factor analyses of the IL knowledge test show no interpretable factor solutions, and thereby provide little evidence of the unidimensionality of these components, either as a group or individually (with one or several identifiable factors respectively). Contrary to some previous factor analytic investigations of freely available IL tests (Leichner et al., 2013; Morley, 2014), this cannot be dismissed on the grounds of insufficient sample sizes. Since the assumption of unidimensionality is not fulfilled, it is not appropriate to use either Cronbach's alpha or other measures of internal consistency for the IL knowledge test.

These findings thereby parallel Beile O'Neil's (2005, p.51) conclusions in her study of the broader IL construct. With no clear factor structure, test items should not be expected to exhibit cohesion. Students can possess knowledge of some concepts of information seeking without having an understanding of other concepts. There is nothing inherent about, for instance, knowing how to use the Boolean 'OR' and knowing how to choose appropriate databases or keywords for your search. Consequently, IL tests, to the extent that they measure heterogeneous constructs, should be considered indexes, rather than scales (cf. DeVelis, 2017; Streiner, 2003), and their reliability should not be assessed using internal consistency measures. One advantage of considering the test an index rather than a scale is that items measure different aspects of IL knowledge and can thus provide valuable formative assessment information over and above the total score on the test.

4.4 Limitations

The four items (18-21) not tested in the pilot may not meet the requirements for item difficulty and discrimination used in item selection. While post hoc testing indicates that all four are within the appropriate range of difficulty for undergraduates, items 18 and 19 may not contribute sufficiently in distinguishing between students with different levels of IL ($r < .30$).

Although reliability and validity for the IL knowledge test and assignment-based measures have been assessed in this study, they should not be considered fully validated for general use, as reliability and validity evidence ideally are established for each specific sample and circumstance (American Educational Research Association, 2014; Streiner & Kottner, 2014).

4.5 Recommended uses of the IL measures

The three measures in this study are each suitable for the aggregated assessment of groups of students for research purposes, and to assess changes in IL levels as a result of interventions. In the latter case, this assumes that the content of the interventions is relevant to information seeking, and source evaluation and use. They may also serve as support for formative assessment and as ingredients in learning activities. To achieve a more accurate evaluation of individual students, for instance for purposes of high stakes educational assessment, we recommend combining several IL measures.

The IL knowledge test is designed to measure undergraduates' knowledge of three core aspects of IL – information seeking, source evaluation and source use. Its scope is therefore narrower than for instance the Open Test of Information Literacy (Hollis et al., 2019) and the most comprehensive commercial tests, yet broader than studies focusing on only one aspect of IL, such as Catalano (2015) and Ondrusek et al. (2005).

4.6 Suggestions for future research

An idea for future research is to follow students longitudinally to assess changes in the association between knowing and doing as they progress. Although several studies have shown increases in IL knowledge *or* skills after interventions, changes in both and correlations between the two over time, measuring with authentic assignments, has not previously been documented. More varied assessment in future research, including both qualitative and additional quantitative measures, could potentially provide stronger evidence for connections between know and do than were found in the present study. We invite others to develop these measures further by following our method, and to produce new measures for other aspects of IL knowledge and practice, thereby expanding this suite of tools in valid and reliable ways.

5. Conclusion

The assessment measures developed and tested in the present study provide a unique combination of tools for quantitatively measuring not only knowledge of core facets of IL, but also applied skills. This does not necessarily mean that these are *complete* measures of students' knowledge and ability to find, evaluate and use sources, but that they measure at least some aspects of these core features of IL that are deemed important in a typical HE context.

We believe this study makes four important contributions. Firstly, although other measures focusing on the same core components of IL are available, few are suitable for multiple disciplines in HE and have been tested with as large a sample size. Secondly, the measures are thoroughly evaluated – also for temporal consistency – allowing reliable and valid interpretations of their results. Thirdly, the study is the first of its kind, as far as we know, to compare *knowing* and *doing* regarding the evaluation and use of sources, utilizing authentic coursework. Fourthly, based on evidence that IL is a heterogeneous construct, the study examines whether IL's source-components can be considered unidimensional constructs on their own or together, and finds little evidence of their homogeneity. We argue, therefore, that unless there is evidence of clear dimensionality, IL tests are best considered indexes rather than scales.

References

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Library Association (1989). *Presidential committee on information literacy: Final report*. Association of College & Research Libraries. http://www.ala.org/acrl/publications/whitepapers/presidential

Association of College and Research Libraries. (2000). *Information literacy competency standards for higher education*. American Library Association. http://www.ala.org/acrl/standards/informationliteracycompetency

Association of College and Research Libraries. (2015). *Framework for information literacy for higher education*. American Library Association. http://www.ala.org/acrl/files/standards/standards.pdf

Beile O'Neil, P. (2005). *Development And validation of the Beile test of information literacy for education (b-tiled).* [Doctoral dissertation, University of Central Florida]. STARS repository. https://stars.library.ucf.edu/etd/530/

Beile, P. (2005). *Development and validation of the information literacy assessment scale for education (ILAS-ED)*. [Conference paper]. American Educational Research Association Annual Meeting, Montreal, Canada. http://eprints.rclis.org/16972/

Bruce, C., Edwards, S., & Lupton, M. (2006). Six frames for information literacy education: A conceptual framework for interpreting the relationships between theory and practice. *Innovation in Teaching and Learning in Information and Computer Sciences, 5(1),* 1–18. https://doi.org/10.11120/ital.2006.05010002

Bujang, M. A., & Baharum, N. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *Archives of Orofacial Sciences, 12(1)*, 1–11.

Bundy, A. (2004). *Australian and New Zealand information literacy framework: Principles, standards and practice* (2nd ed.). Australian and New Zealand Institute for Information Literacy. http://www.libnet.sh.cn/upload/htmleditor/File/130620025617.pdf

Catalano, A. (2015). The effect of a situated learning environment in a distance education information literacy course. *The Journal of Academic Librarianship, 41(5),* 653–659. https://doi.org/10.1016/j.acalib.2015.06.008

CILIP Information literacy group. (2018). CILIP definition of information literacy 2018. CILIP: The Library and Information Association. https://www.cilip.org.uk/news/421972/What-is-information-literacy.htm

Coonan, E., & Secker, J. (2011). A new curriculum for information literacy (ANCIL): Curriculum and supporting documents. University of Cambridge. https://www.repository.cam.ac.uk/handle/1810/244638

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52(4),* 281. https://doi.org/10.1037/h0040957

DeMars, C. E. (2017). Classical test theory and item response theory. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vol. 1-2, pp. 49–73). John Wiley & Sons.

DeVelis, R. F. (2017). Scale development: Theory and applications (4th ed.). Sage.

Gross, M., & Latham, D. (2007). Attaining information literacy: An investigation of the relationship between skill level, self-estimates of skill, and library anxiety. *Library and Information Science Research, 29*(3), 332–353. https://doi.org/10.1016/j.lisr.2007.04.012

Gross, M., & Latham, D. (2012). What's skill got to do with it? Information literacy skills and self-views of ability among first-year college students. *Journal of the American Society for Information Science and Technology, 63(3),* 574–583. https://doi.org/10.1002/asi.21681

Gross, M., & Latham, D. (2013). Addressing below proficient information literacy skills: Evaluating the efficacy of an evidence-based educational intervention. *Library and Information Science Research, 35(3),* 181–190. https://doi.org/10.1016/j.lisr.2013.03.001

Gross, M., Latham, D., & Julien, H. (2018). What the framework means to me: Attitudes of academic librarians toward the ACRL framework for information literacy for higher education. *Library & Information Science Research, 40(3)*, 262–268. https://doi.org/10.1016/j.lisr.2018.09.008

Hollis, H. (2018). Information literacy as a measurable construct: A need for more freely available, validated and wide ranging instruments. *Journal of Information Literacy, 12(2), 76*–88. https://doi.org/10.11645/12.2.2409

Hollis, H., Rachitskiy, M., & van der Leer, L. (2019). The development and face validity testing of the Open Test of Information Literacy with context-specific add-ons: OTIL. *LIBER Quarterly, 29(1).* http://doi.org/10.18352/lq.10264

Hughes, D. J. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp.751–779). John Wiley & Sons.

Ivanitskaya, L., O'Boyle, I., & Casey, A. M. (2006). Health information literacy and competencies of information age students: Results from the interactive online Research Readiness Self-Assessment (RRSA). *Journal of Medical Internet Research, 8(2).*

Julien, H., & Hoffman, C. (2008). Information literacy training in Canada's public libraries. *The Library Quarterly, 78(1),* 19–41. https://doi.org/10.1086/523908

Leichner, N., Peter, J., Mayer, A. K., & Krampen, G. (2014). Assessing information literacy programmes using information search tasks. *Journal of Information Literacy, 8(1),* 3–20. https://doi.org/10.11645/8.1.1870

Mahmood, K. (2017). A systematic review of evidence on psychometric properties of information literacy tests. *Library Review, 66(6/7),* 442–455. https://doi.org/10.1108/LR-02-2017-0015

Morley, S. K. (2014). *Initial development of a medical information literacy questionnaire* [Doctoral dissertation, The University of New Mexico College of Education, USA]. The University of New Mexico repository. https://digitalrepository.unm.edu/educ_ifce_etds/30/

Mulaik, S. A. (2009). *Foundations of factor analysis* (2nd ed.). CRC Press.

Multon, K. D. (2012). Interrater reliability. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp.627–628). https://dx.doi.org/10.4135/9781412961288

Nierenberg, E., & Fjeldbu, Ø. G. (2015). How much do first-year undergraduate students in Norway know about information literacy? *Journal of Information Literacy, 9(1),* 15–33. https://doi.org/10.11645/9.1.1983

Nierenberg, E., Låg, T., & Dahl, T. I. (2021*). Replication Data for: Knowing and doing: The development of information literacy measures to assess knowledge and practice* (DataverseNO: Version V2) [Data set, suvey and evaluation measures]. DataverseNO. https://doi.org/doi:10.18710/L60VDI

Oakleaf, M. (2008). Dangers and opportunities: A conceptual map of information literacy approaches. *portal: Libraries and the Academy, 8(3),* 233–253. https://doi.org/10.1353/pla.0.0011

Oakleaf, M. (2014). A roadmap for assessing student learning using the new framework for information literacy for higher education. *The Journal of Academic Librarianship, 5(40),* 510–514. https://doi.org/10.1016/j.acalib.2014.08.001

Oakleaf, M., Millet, M. S., & Kraus, L. (2011). All together now: Getting faculty, administrators, and staff engaged in information literacy assessment. *portal: Libraries and the Academy, 11(3),* 831–852. https://doi.org/10.1353/pla.2011.0035

Ondrusek, A., Dent, V. F., Bonadie-Joseph, I., & Williams, C. (2005). A longitudinal study of the development and evaluation of an information literacy test. *Reference Services Review, 33(4),* 388–417. https://doi.org/10.1108/00907320510631544

Podgornik, B. B., Dolničar, D., Šorgo, A., & Bartol, T. (2015). Development, testing, and validation of an information literacy test (ILT) for higher education. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.23586

Polkinghorne, S., & Wilton, S. (2010). Research is a verb: Exploring a new linformation literacy-embedded undergraduate research methods course. *Canadian Journal of Information and Library Science, 34(4),* 457–473. https://doi.org/10.1353/ils.2010.0008

Schilling, K., & Applegate, R. (2007). *Evaluating library instruction: Measures for assessing educational quality and impact*. [Conference paper]. The Thirteenth National Conference of the Association of College and Research Libraries, Baltimore, MD, United States. http://www.ala.org/acrl/conferences/confsandpreconfs/national/baltimore/baltimore

Schilling, K., & Applegate, R. (2012). Best methods for evaluating educational impact: A comparison of the efficacy of commonly used measures of library instruction. *Journal of the Medical Library Association, 100(4),* 258–269. https://doi.org/10.3163/1536-5050.100.4.007

SCONUL Working Group on Information Literacy. (2011). *The SCONUL seven pillars of information literacy: Core model for higher education*. SCONUL. https://www.sconul.ac.uk/sites/default/files/documents/coremodel.pdf

Smith, J. K., Given, L. M., Julien, H., Ouellette, D., & DeLong, K. (2013). Information literacy proficiency: Assessing the gap in high school students' readiness for undergraduate academic work. *Library & Information Science Research, 35(2),* 88–96. https://doi.org/10.1016/j.lisr.2012.12.001

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80(3),* 217–222. https://doi.org/10.1207/S15327752JPA8003_01

Streiner, D. L., & Kottner, J. (2014). Recommendations for reporting the results of studies of instrument and scale development and testing. *Journal of Advanced Nursing, 70(9),* 1970–1979. https://doi.org/10.1111/jan.12402

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56(2),* 197–208. https://doi.org/10.1177/0013164496056002001

Thompson, S. (2003). *Information literacy meeting of experts: Conference report*. [Conference paper]. Information Literacy Meeting of Experts, Prague, Czech Republic. http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/themes/info_lit_meeting_prague_2003.pdf

Thorsen, S. V., & Bjorner, J. B. (2009). Reliability of the Copenhagen Psychosocial Questionnaire. *Scandinavian Journal of Public Health, 38(3_suppl),* 25–32. https://doi.org/10.1177/1403494809349859

Walsh, A. (2009). Information literacy assessment: Where do we start? *Journal of Librarianship and Information Science, 41(1),* 19–28. https://doi.org/10.1177/0961000608099896

Walton, G., & Hepworth, M. (2012). Using assignment data to analyse a blended information literacy intervention: A quantitative approach. *Journal of Librarianship and Information Science, 45(1),* 53–63. https://doi.org/10.1177/0961000611434999

Appendix

Appendix A

Exploratory factor analyses

Contents:

1. Exploratory factor analysis on the IL test items – Pilot sample (17 items)

Raw data and correlation matrices used in these analyses are available Nierenberg, Låg, and Dahl, 2021.

After excluding items on the basis of low item-total correlations and/or extreme P-values, we performed exploratory factor analyses (EFA) on the 17 remaining items using the data from the Pilot sample. The aim was to explore the dimensionality of the IL test, if it has any, and possibly to further reduce the number of items.

Since items in the IL test are scored dichotomously as correct or incorrect, we used a matrix of tetrachoric correlations among items as the basis for the EFA.

All factor analyses were done using the psych package for R (Revelle, 2020).

1.1 Data suitability for FA
With a sample size of 268 observations, the observation-to-item ratio is 15.76 for these analyses.

We used a number of different criteria (see e.g., Dziuban and Shirkey, 1974) to assess the matrix' suitability for FA. Visual inspection of the matrix revealed few sizeable correlations (only 33 of 136 or 24% were stronger than .3, and only 1 was stronger than .5). The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .56, and hardly acceptable. The Bartlett test of sphericity was significant ($\chi^2(136) = 1389.4$, $p < .001$). Communalities were generally low.

With the exception of a fairly good observation-to-item ratio and a significant Bartlett test (although the latter is not a very reliable indicator), the other indicators seem to cast some doubt on the factorability of these data. We nevertheless decided to perform them, keeping in mind that any sensible solution would need to be cautiously endorsed and preferably replicated.

1.2 Extraction and rotation methods

Since items in the IL test are scored dichotomously, multivariate normality is violated. Therefore, the principal axis factoring method was chosen (cf. Fabrigar, Wegener, MacCallum & Strahan, 1999). We had no strong prior assumptions regarding the intercorrelatedness of potential factors, and decided to allow solutions with correlated factors, using an oblique (the Promax with Kaiser normalization) rotation.

1.3 Number of factors

The 17 items remaining after the selection process described in the main manuscript were all related to three core aspects of IL. One might, therefore, expect the most interpretable factor **solution** to be a 3-factor solution. On the other hand, if IL is conceived of as "one underlying variable" or "singular construct" (Hollis, 2018, s. 76-77), a latent attribute that causes the observable behaviours that IL tests measure, then one would expect a strong first factor, with remaining factors accounting for very little variance (Streiner, 2003). A third possibility is that a relatively broadly conceived source oriented IL test measures a number of rather different facets that may or may not be related to each other in any other way than that they tend to be seen as part of IL, in which case we should not expect any clear or meaningful factor structure.

Given that all of these are reasonable a priori possibilities, we hoped in addition to rely on initial eigenvalues in deciding the number of factors to extract. Initial eigenvalues and cumulative percentage of variance explained are reported in Table A1. Figure A1 displays the scree plot.

Table A1: Initial eigenvalues and percentage of variance explained for factors extracted from pilot data on 17 items from the IL test

Factor	Eigenvalue	Cumulative % variance
1	4.26	26.6
2	1.41	35.5
3	1.31	43.7
4	1.22	51.3
5	1.13	58.3
6	1.05	64.8

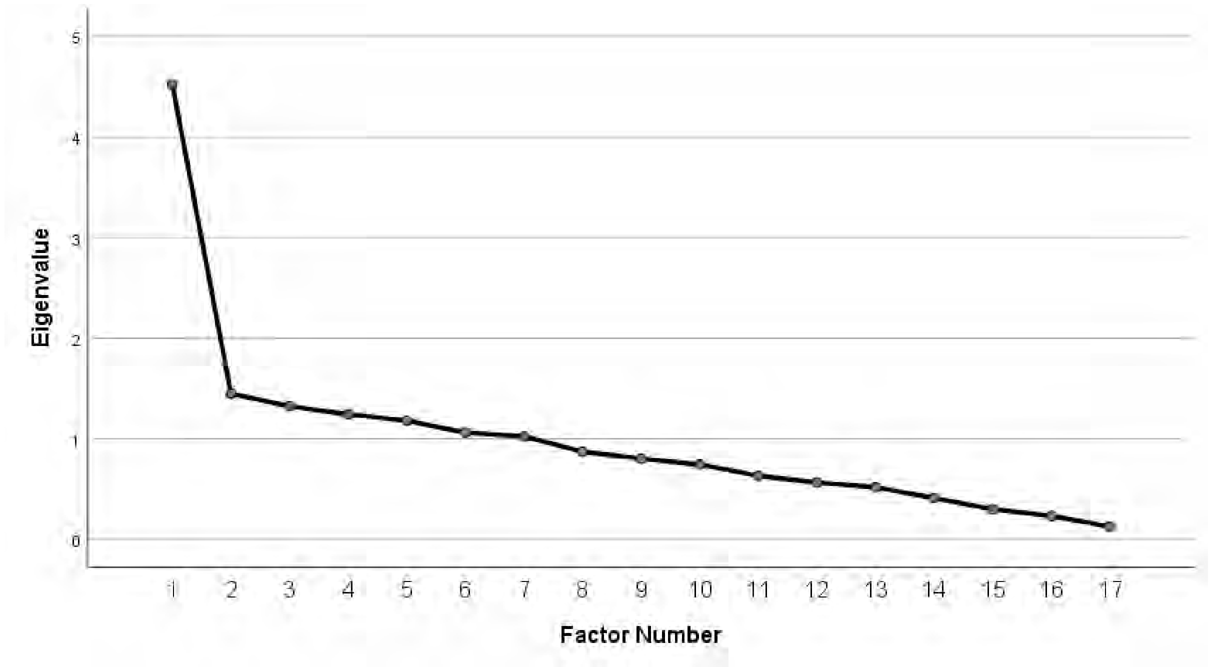

Figure A1: Scree plot of initial eigenvalues for factors extracted from pilot data on 17 items from the IL test

The scree plot seems to display only one clear "elbow", indicating that a one-factor solution may be preferable. However, the first factor explains only 26.6% of total variance. And, while subsequent factors increase explained variance only marginally, it still seemed sensible to explore several different solutions.

We explored solutions with everything from 1 to 5 factors. Results for one-, two- and three-factor solutions, which seemed to us marginally more sensible than the other solutions, are reported in the subsections below.

One-factor solution. Factor loadings for the one-factor solution are reported in Table A2.

Table A2: Factor loadings for a one-factor solution from pilot data on 17 items from the IL test

	Item	**Factor 1**	**IL aspect***	**Communalities⁺**
14	What kind of information does not need to be cited?	.59	Use	.35
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	.58	Evaluate	.34
5	"The liberal world order will continue…" (…) Would you say this quote is (…)	.54	Evaluate	.29
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	.52	Evaluate	.27
8	Which reference style should you use in your article if this is not (…)	.52	Use	.27
13	Is it allowed to use an image from a webpage in your article?	.51	Use	.26
10	In which of the sentences below you do not need to cite the source?	.50	Use	.25
3	Which source is the least suitable for finding information for your article on…	.50	Evaluate	.25
4	Which of the following is not necessary in order to critically evaluate a source?	.48	Evaluate	.23
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	.47	Use	.22
1	What characterizes a scholarly article?	.47	Evaluate	.22
2	You find these two books (…). Which of these statements is most correct?	.44	Evaluate	.19
17	You get only two results from (…) What should you do to get more?	.42	Find	.18
9	What is the most important reason to use sources when writing a paper?	.40	Use	.16
15	You are writing a paper on (…) Which search gives the best results?	.39	Find	.15
12	[Example shown.] Which source should be cited?	.34	Use	.12
16	The most effective way to follow the newest research on (…)	.16	Find	.03

Notes. Principal axis factor extraction; RMSEA = 0.14 [.13, .15]; TLI = .45
*These are aspects that the items were originally generated on the basis of
⁺ Extracted

Two-factor solution. Factor loadings for the two-factor solution are reported in Table A3.

Table A3: Factor loadings for a 2-factor solution from pilot data on 17 items from the IL test

	Item	**Factor 1**	**Factor 2**	**IL aspect***	**Communalities⁺**
14	What kind of information does not need to be cited?	**1.16**	-.43	Use	.86
13	Is it allowed to use an image from a webpage in your article?	**.56**	.00	Use	.32
2	You find these two books (…). Which of these statements is most correct?	**.55**	-.07	Evaluate	.26
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	**.40**	.12	Use	.24
17	You get only two results from (…) What should you do to get more?	.31	.14	Find	.18
9	What is the most important reason to use sources when writing a paper?	.31	.13	Use	.17
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	-.04	**.67**	Evaluate	.42
12	[Example shown.] Which source should be cited?	-.23	**.60**	Use	.23
10	In which of the sentences below do you not need to cite the source?	.02	**.52**	Use	.29
1	What characterizes a scholarly article?	-.01	**.51**	Evaluate	.26
3	Which source is the least suitable for finding information for your article on…	.07	**.47**	Evaluate	.27
4	Which of the following is not necessary in order to critically evaluate a source?	.07	**.45**	Evaluate	.25
15	You are writing a paper on (…) Which search gives the best results?	-.01	**.43**	Find	.18
8	Which reference style should you use in your article if this is not (…)	.20	.36	Use	.27
5	"The liberal world order will continue…" (…) Would you say this quote is (…)	.26	.32	Evaluate	.28
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	.25	.31	Evaluate	.27
16	The most effective way to follow the newest research on (…)	.09	.09	Find	.03

Notes. Principal axis factor extraction; Promax rotation with Kaiser normalization; Factor loadings larger than .40 are in bold; The correlation between the two factors was .68; RMSEA = .14 [.13, .15]; TLI = .46

* These are aspects that the items were originally generated on the basis of.

⁺ Extracted

Three-factor solution. Factor loadings for the three-factor solution are reported in Table A4. Correlations between factors are reported in Table A5.

Table A4: Factor loadings for a 3-factor solution from pilot data on 17 items from the IL test

	Item	Factor 1	Factor 2	Factor 3	IL aspect*	Communalities⁺
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	**.78**	.00	-.22	Evaluate	.51
10	In which of the sentences below you do not need to cite the source?	**.67**	.08	-.32	Use	.42
12	[Example shown.] Which source should be cited?	**.55**	-.21	.04	Use	.21
3	Which source is the least suitable for finding information for your article on…	**.46**	.09	-.02	Evaluate	.27
1	What characterizes a scholarly article?	**.45**	-.05	.19	Evaluate	.28
4	Which of the following is not necessary in order to critically evaluate a source?	**.42**	.06	.07	Evaluate	.25
8	Which reference style should you use in your article if this is not (…)	.32	.17	.11	Use	.26
15	You are writing a paper on (…) Which search gives the best results?	.38	-.04	.14	Find	.19
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	.27	.22	.12	Evaluate	.27
14	What kind of information does not need to be cited?	-.40	**1.24**	-.13	Use	.97
13	Is it allowed to use an image from a webpage in your article?	.04	**.54**	-.04	Use	.31
2	You find these two books (…). Which of these statements is most correct?	-.12	**.48**	.21	Evaluate	.28
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	.10	.36	.10	Use	.23
9	What is the most important reason to use sources when writing a paper?	.19	.33	-.13	Use	.19
17	You get only two results from (…) What should you do to get more?	.15	.31	-.01	Find	.18
16	The most effective way to follow the newest research on (…)	-.11	-.06	**.61**	Find	.30
5	"The liberal world order will continue…" (…) Would you say this quote is (…)	.19	.15	**.45**	Evaluate	.38

Notes. Principal axis factor extraction; Promax rotation with Kaiser normalization; Factor loadings larger than .40 are in bold; RMSEA = .14 [.13, .15]; TLI = .43

*These are aspects that the items were originally generated on the basis of.

⁺Extracted

Table A5: Factor correlation matrix

Factor	1	2	3
1		.66	.42
2			.44
3			

1.4 Evaluation of factor solutions from pilot sample data

One notable feature of these results is that none of the factors in the two- or three-factor solutions seem to align sensibly to the core IL aspects from which test items were derived, nor indeed to any other meaningful pattern. Neither does there seem to be any support for endorsing a single-factor solution, partly because of the very modest proportion of variance explained. Overall, the items seem to have mostly small correlations among each other, and this is reflected in generally low factor loadings in all the solutions. Since no very clear factor structure is evident, and because we consider broad coverage of the three source-related core IL aspects paramount, we found no basis in these results for further pruning of the test.

2. Exploratory factor analysis on the IL test items – Pretest sample (17 items)

As described in the main manuscript, a revised 21-item version (of which 17 items were identical to the those administered to the Pilot sample) of the IL test was administered to a new sample (*N* = 260) of undergraduate students (the Pretest sample).

Although we were unable to make sense of the factor solutions on the Pilot sample data, we were curious as to whether the solutions from the Pilot sample would reproduce in a separate, similar sample. We wanted to remain open-minded as to whether the solutions produced by the analyses on the Pilot sample data were just artefacts of sample idiosyncrasies, or perhaps in fact systematically related to (some unobvious) properties of the test items themselves. If the solutions reproduced in a separate sample, they would perhaps merit a closer look.

At this point we would like to emphasize the exploratory intent behind these analyses. As stated in the main manuscript and in the first section of this supplement, we had few strong reasons to expect a particular factor structure for the IL test to be more likely to emerge than others. Thus, EFA is clearly the most appropriate analytic approach.

However, since the goal of exploratory factor analysis is to infer the likely structure of an instrument when used in a population, it is important to assess whether the same basic factor structure replicates in a separate sample from that population (Osborne, 2014; Osborne & Fitzpatrick, 2012). Factor solutions are notoriously hard to reproduce, even when there is an adequate sample size and a relatively clear factor structure, although both sample size and structure clarity are likely to increase the probability of reproducing the solution.

The analyses presented here were performed on a matrix of tetrachoric correlations (based on the same 17 items analyzed in the Pilot sample).

2.1 Data suitability, extraction and rotation methods

With a sample size of 260 observations, the observation-to-item ratio is 15.29 for these analyses.

Visual inspection of the matrix again revealed few sizeable correlations (only 19 of 136 or 14% were stronger than .3, and none were stronger than .5). The KMO measure of sampling adequacy was .57, but the Bartlett test of sphericity was significant ($\chi^2(136) = 869.6$, $p < .001$). Communalities were generally low. Again, the factorability of these data is not very good.

The same general analysis approach (principal axis factoring and Promax rotation with Kaiser normalization) was used to arrive at the factor solutions presented here.

Initial eigenvalues and cumulative percentage of variance explained are reported in Table A6. Figure A2 displays the scree plot.

Table A6: Initial eigenvalues and percentage of variance explained for factors extracted from Pretest sample data on 17 items from the IL test

Factor	Eigenvalue	Cumulative % variance
1	3.29	19.3
2	1.71	29.4
3	1.38	37.5
4	1.32	45.3
5	1.23	52.5
6	1.15	49.2

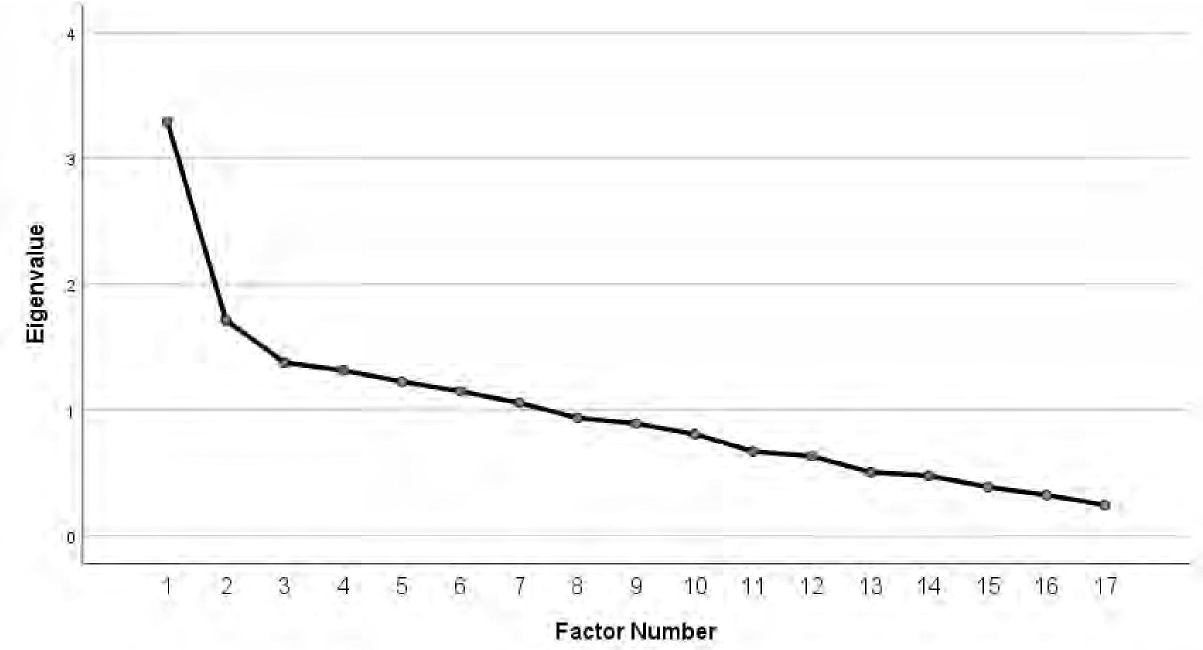

Figure A2: Scree plot of initial eigenvalues for factors extracted from the Pretest sample data on 17 items from the IL test

2.2 Factor solutions

One-factor solution. Factor loadings for the one-factor solution are reported in Table A7.

Table A7: Factor loadings for a one-factor solution using data from the Pretest sample on 17 items from the IL test.

	Item	Factor 1	IL aspect*	Communalities⁺
13	Is it allowed to use an image from a webpage in your article?	.62	Use	.38
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	.59	Evaluate	.35
2	You find these two books (…). Which of these statements is most correct?	.54	Evaluate	.29
17	You get only two results from (…) What should you do to get more?	.50	Find	.25
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	.47	Evaluate	.22
3	Which source is the least suitable for finding information for your article on…	.46	Evaluate	.21
14	What kind of information does not need to be cited?	.40	Use	.16
5	“The liberal world order will continue…” (…) Would you say this quote is (…)	.35	Evaluate	.12
8	Which reference style should you use in your article if this is not (…)	.35	Use	.12
15	You are writing a paper on (…) Which search gives the best results?	.33	Find	.11
1	What characterizes a scholarly article?	.28	Evaluate	.08
9	What is the most important reason to use sources when writing a paper?	.27	Use	.07
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	.23	Use	.05
12	[Example shown.] Which source should be cited?	.22	Use	.05
16	The most effective way to follow the newest research on (…)	.16	Find	.03
10	In which of the sentences below you do not need to cite the source?	.14	Use	.02
4	Which of the following is not necessary in order to critically evaluate a source?	.12	Evaluate	.01

Notes. Principal axis factor extraction; RMSAE = .11 [.10, .12]; TLI = .39
*These are aspects that the items were originally generated on the basis of
⁺ Extracted

Two-factor solution. Factor loadings for the two-factor solution are reported in Table A8.

Table A8: Factor loadings (pattern matrix) for a 2-factor solution from Pretest sample data on 17 items from the IL test

	Item	**Factor 1**	**Factor 2**	**IL aspect***	**Communalities⁺**
2	You find these two books (…). Which of these statements is most correct?	**.65**	-.02	Evaluate	.41
13	Is it allowed to use an image from a webpage in your article?	**.62**	.10	Use	.44
3	Which source is the least suitable for finding information for your article on…	**.53**	-00	Evaluate	.28
15	You are writing a paper on (…) Which search gives the best results?	**.41**	-.04	Find	.16
17	You get only two results from (…) What should you do to get more?	**.41**	.18	Find	.25
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	.36	-.11	Use	.11
14	What kind of information does not need to be cited?	.35	.12	Use	.17
16	The most effective way to follow the newest research on (…)	.21	-.03	Find	.04
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	.08	**.55**	Evaluate	.34
5	“The liberal world order will continue…” (…) Would you say this quote is (…)	.00	**.48**	Evaluate	.23
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	.29	**.44**	Evaluate	.37
4	Which of the following is not necessary in order to critically evaluate a source?	-.23	**.43**	Evaluate	.17
12	[Example shown.] Which source should be cited?	-.07	.39	Use	.14
8	Which reference style should you use in your article if this is not (…)	.09	.36	Use	.16
1	What characterizes a scholarly article?	.02	.34	Evaluate	.12
9	What is the most important reason to use sources when writing a paper?	.09	.24	Use	.08
10	In which of the sentences below you do not need to cite the source?	-.03	.22	Use	.05

Notes. Principal axis factor extraction; Promax rotation with Kaiser normalization; Factor loadings larger than .40 are in bold; The correlation between the two factors was .36; RMSEA = .11 [.10, .12]; TLI = .43

* These are aspects that the items were originally generated on the basis of.

⁺ Extracted

Three-factor solution. Factor loadings for the three-factor solution are reported in Table A9. Correlations between factors are reported in Table A10.

Table A9: Factor loadings (pattern matrix) for a 3-factor solution from Pretest sample data on 17 items from the IL test

	Item	**Factor 1**	**Factor 2**	**Factor 3**	**IL aspect***	**Communalities⁺**
2	You find these two books (…). Which of these statements is most correct?	**.80**	-.17	-.03	Evaluate	.56
3	Which source is the least suitable for finding information for your article on…	**.52**	-.01	.02	Evaluate	.26
13	Is it allowed to use an image from a webpage in your article?	**.49**	.21	.26	Use	.41
14	What kind of information does not need to be cited?	**.40**	.06	-.01	Use	.18
17	You get only two results from (…) What should you do to get more?	**.40**	.17	.04	Find	.24
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	.34	-.10	.12	Use	.12
16	The most effective way to follow the newest research on (…)	.23	-.05	-.02	Find	.05
4	Which of the following is not necessary in order to critically evaluate a source?	-.31	**.53**	.07	Evaluate	.24
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	.13	**.49**	-.06	Evaluate	.32
12	[Example shown.] Which source should be cited?	-.10	**.42**	-.01	Use	.16
8	Which reference style should you use in your article if this is not (…)	.04	**.42**	.11	Use	.19
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	.32	.39	-.06	Evaluate	.37
9	What is the most important reason to use sources when writing a paper?	.06	.28	.05	Use	.09
10	In which of the sentences below you do not need to cite the source?	-.04	.24	.04	Use	.05
1	What characterizes a scholarly article?	.12	.24	-.15	Evaluate	.13
15	You are writing a paper on (…) Which search gives the best results?	.24	.26	**.97**	Find	1.01
5	"The liberal world order will continue…" (…) Would you say this quote is (…)	.12	.35	-.34	Evaluate	.36

Notes. Principal axis factor extraction; Promax rotation with Kaiser normalization; Factor loadings larger than .40 are in bold; RMSEA = .10 [.09, .12]; TLI = .48

*These are aspects that the items were originally generated on the basis of.

⁺Extracted

Table A10: Factor correlation matrix

Factor	1	2	3
1		.40	-.05
2			-.14
3			

2.3 Evaluation of factor solutions from Pretest sample data and comparison to solutions from Pilot sample data

As with the pilot sample data, there is no readily interpretable factor solution, which may simply be due to the fact that the correlation matrix is not very suitable for factorization. Comparing the factor solutions from the two samples, one would be hard pressed to say that the few similarities reach a minimum replicability threshold of the same basic factor structure. For instance, comparing the two-factor solutions, several items load on different factors (e.g., item 15 and 3 are grouped with items 2, 14, 13, and 11 in the Pretest sample solution, but not in the Pilot sample solution). Even when items group on the same factor in solutions from the two samples, loading magnitudes are somewhat different.

In sum, when considering explored solutions from the Pilot sample data and the Pretest sample data, there seem to be no readily interpretable solutions, and the solutions emerging from the two samples do not match very well. No clear dimensionality is evident for the IL test based on these analyses.

3. Exploratory factor analysis on the complete IL test – Undergraduate and graduate samples (21 items)

In spite of the fact that the observation-to-item ratios of the analyses described in the previous section are fairly good compared to what seems to be common practice in fields that employ factor analysis (cf. for instance Ford, MacCallum & Tait, 1986, who found that 56% of analyses published over a 10-year period had observation-to-item ratios of less than 10:1), they are still slightly short of ideals recommended by some (e.g., the 500 observation limit for considering a sample size "very good" according to Comrey & Lee, 1992, or the 400 observation minimum suggested by Aleamoni, 1976).

Furthermore, although no readily interpretable structure emerged from the analyses described in the previous sections, we deemed it possible that one might emerge from i) a larger sample; ii) a sample including more experienced students likely to manifest higher levels of IL; iii) a matrix based on the full 21-item test, including revised items.

We therefore merged the Pretest sample (*N* = 260) with the Graduate student sample (*N* = 366) and again explored the possible dimensionality of the IL test using EFA.

3.1 Data suitability for FA

With a total sample size of 626 observations, the observation-to-item ratio is 29.81 for these analyses.

Visual inspection of the matrix revealed few sizeable correlations (only 39 of 210 or 18.6% were stronger than .3, and only 3 of 210 or 1.4% were stronger than .5). The KMO measure of sampling adequacy was .76, which is typically considered adequate. The Bartlett test of sphericity was significant ($\chi^2(210) = 3668.4$, $p < .001$). Communalities, however, were generally low.

Again, despite a conventionally adequate KMO-value, the limited number of intercorrelations among the items indicates relatively poor factorability. We nevertheless decided to perform the analyses, again keeping in mind that any sensible solution would need to be cautiously endorsed and preferably replicated.

3.2 Extraction and rotation methods

The considerations regarding extraction and rotation methods from the analyses described above were applied to the current sample.

3.3 Number of factors

Again, given the varied reasonable a priori possibilities outlined in the first section of this supplement, we hoped in addition to rely on initial eigenvalues in deciding the number of factors to extract. Initial eigenvalues and cumulative percentage of variance explained are reported in Table A11. Figure A3 displays the scree plot.

Table A11: Initial eigenvalues and percentage of variance explained for factors extracted from data on the complete 21 item IL test.

Factor	Eigenvalue	Cumulative % variance
1	5.25	25.0
2	1.53	32.2
3	1.39	38.9
4	1.32	45.1
5	1.21	50.9
6	1.10	56.1

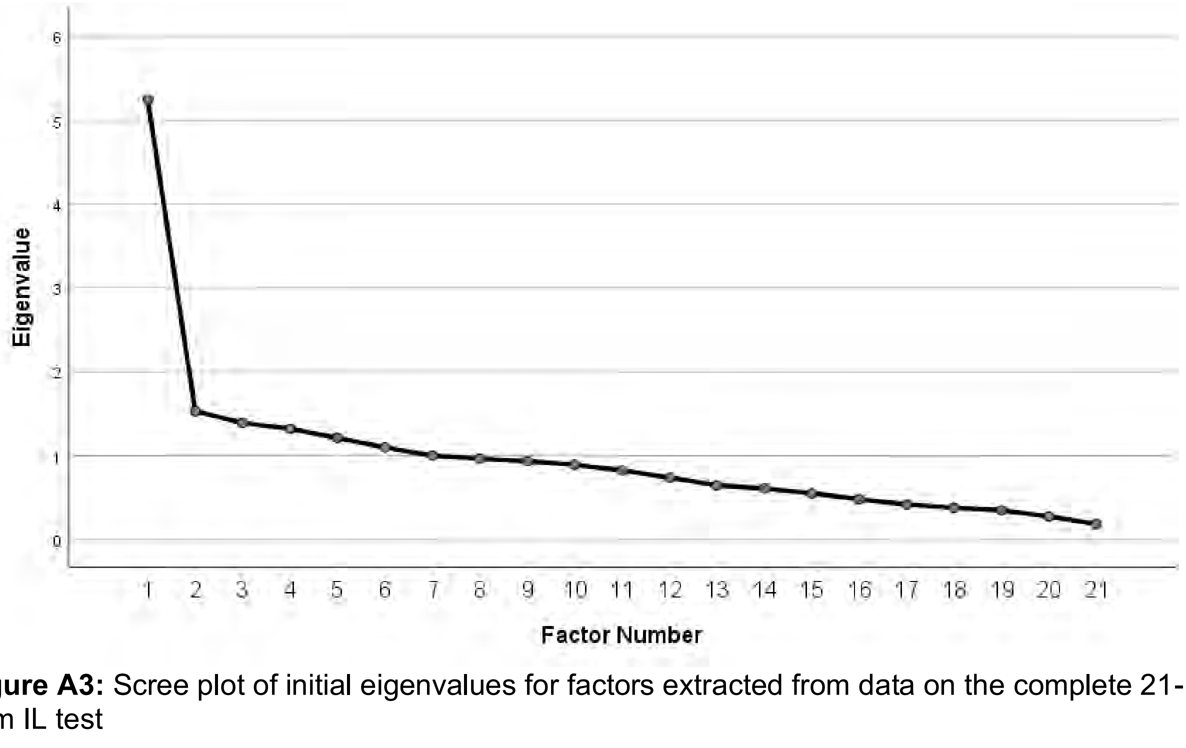

Figure A3: Scree plot of initial eigenvalues for factors extracted from data on the complete 21-item IL test

The scree plot seems to display only one clear "elbow", indicating that a one-factor solution may be preferable. However, the first factor explains only 25% of the total variance. And, while subsequent factors increase explained variance only marginally, it still seemed sensible to explore several different solutions.

We explored solutions with everything from 1 to 5 factors. Results for one-, two- and three-factor solutions are reported in the subsections below.

One-factor solution. Factor loadings for the one-factor solution are presented in Table A12.

Table A12: Factor loadings for a one-factor solution using data from the complete 21-item IL test

	Item	**Factor 1**	**IL aspect***	**Communalities⁺**
13	Is it allowed to use an image from a webpage in your article?	.72	Use	.52
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	.67	Evaluate	.45
8	Which reference style should you use in your article if this is not (…)	.63	Use	.39
20	Which statement is correct? [Statements about searching.]	.62	Find	.39
21	(…) If you do not get sufficient results the first time, what should you do?	.61	Find	.37
15	You are writing a paper on (…) Which search gives the best results?	.60	Find	.36
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	.55	Evaluate	.30
1	What characterizes a scholarly article?	.48	Evaluate	.23
3	Which source is the least suitable for finding information for your article on…	.45	Evaluate	.20
5	"The liberal world order will continue…" (…) Would you say this quote is (…)	.44	Evaluate	.20
17	You get only two results from (…) What should you do to get more?	.41	Find	.17
4	Which of the following is not necessary in order to critically evaluate a source?	.40	Evaluate	.16
10	In which of the sentences below you do not need to cite the source?	.37	Use	.14
16	The most effective way to follow the newest research on (…)	.36	Find	.13
2	You find these two books (…). Which of these statements is most correct?	.34	Evaluate	.12
18	If you wrote comp* in the search field of a database such as (…)	.32	Find	.10
14	What kind of information does not need to be cited?	.31	Use	.10
9	What is the most important reason to use sources when writing a paper?	.29	Use	.09
19	In which case is it not recommended to use OR between words (…)	.26	Find	.07
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	.22	Use	.05
12	[Example shown.] Which source should be cited?	.14	Use	.02

Notes. Principal axis factor extraction
*These are aspects that the items were originally generated on the basis of; RMSEA = .11 [.10, .11]; TLI = .57
⁺Extracted

Two-factor solution. Factor loadings for the two-factor solution are reported in Table A13.

Table A13: Factor loadings (pattern matrix) for a two-factor solution using data from the complete 21-item IL test

	Item	Factor 1	Factor2	IL aspect*	Communalities⁺
18	If you wrote comp* in the search field of a database such as (…)	**.83**	-.50	Find	.41
8	Which reference style should you use in your article if this is not (…)	**.67**	.01	Use	.46
15	You are writing a paper on (…) Which search gives the best results?	**.55**	.10	Find	.38
4	Which of the following is not necessary in order to critically evaluate a source?	**.48**	-.06	Evaluate	.20
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	**.45**	.15	Evaluate	.30
21	(…) If you do not get sufficient results the first time, what should you do?	**.44**	.22	Find	.37
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	**.42**	.31	Evaluate	.45
10	In which of the sentences below you do not need to cite the source?	.39	.01	Use	.15
16	The most effective way to follow the newest research on (…)	.37	.02	Find	.15
1	What characterizes a scholarly article?	.35	.17	Evaluate	23
5	"The liberal world order will continue…" (…) Would you say this quote is (…)	.32	.16	Evaluate	.19
19	In which case is it not recommended to use OR between words (…)	.22	.06	Find	.07
12	[Example shown.] Which source should be cited?	.11	.05	Use	.02
3	Which source is the least suitable for finding information for your article on…	-.27	**.83**	Evaluate	.48
13	Is it allowed to use an image from a webpage in your article?	.21	**.61**	Use	.58
20	Which statement is correct? [Statements about searching.]	.20	**.49**	Find	.42
2	You find these two books (…). Which of these statements is most correct?	-.03	**.42**	Evaluate	.17
14	What kind of information does not need to be cited?	.01	.35	Use	.13
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	-.04	.29	Use	.07
17	You get only two results from (…) What should you do to get more?	.19	.27	Find	.17
9	What is the most important reason to use sources when writing a paper?	.11	.21	Use	.09

Notes. Principal axis factor extraction; Promax rotation with Kaiser normalization; Factor loadings larger than .40 are in bold; The correlation between the two factors was .58; RMSEA = .10 [.10, .11]; TLI = .59

* These are aspects that the items were originally generated on the basis of.

⁺ Extracted

Three-factor solution. Factor loadings for the three-factor solution are reported in Table A14. Correlations between factors are reported in Table A15.

Table A14: Factor loadings (pattern matrix) for a 3-factor solution using data from the complete 21-item IL test

	Item	Factor 1	Factor 2	Factor 3	IL aspect*	Communalities $^+$
13	Is it allowed to use an image from a webpage in your article?	**.82**	-.11	.21	Use	.75
20	Which statement is correct? [Statements about searching.]	**.70**	-.08	.20	Find	.51
3	Which source is the least suitable for finding information for your article on…	**.51**	.42	-.39	Evaluate	.50
6	All of these criteria are useful for assessing the reliability of a journal article, (…)	.32	.22	.28	Evaluate	.44
21	(…) If you do not get sufficient results the first time, what should you do?	.25	.19	.30	Find	.37
11	John uses paragraphs from his own essay (…) Is this considered plagiarism?	.34	-.08	.00	Use	.09
2	You find these two books (…). Which of these statements is most correct?	.29	.21	-.10	Evaluate	.16
1	What characterizes a scholarly article?	-.06	**.53**	.10	Evaluate	.30
7	You are writing a paper about Arctic ice melting. Who most likely has the least…	-.02	**.47**	.21	Evaluate	.35
4	Which of the following is not necessary in order to critically evaluate a source?	-.21	**.45**	.26	Evaluate	.27
14	What kind of information does not need to be cited?	.10	**.42**	-.16	Use	.18
17	You get only two results from (…) What should you do to get more?	.14	.30	.05	Find	.18
9	What is the most important reason to use sources when writing a paper?	.01	.38	-.05	Use	.13
5	"The liberal world order will continue…" (…) Would you say this quote is (…)	.06	.31	.16	Evaluate	.21
19	In which case is it not recommended to use OR between words (…)	-.12	.39	.04	Find	.12
12	[Example shown.] Which source should be cited?	.05	.05	.07	Use	.02
18	If you wrote comp* in the search field of a database such as (…)	-.22	-.04	**.71**	Find	.40
15	You are writing a paper on (…) Which search gives the best results?	.29	-.02	**.49**	Find	.44
8	Which reference style should you use in your article if this is not (…)	.01	.33	**.44**	Use	.46
16	The most effective way to follow the newest research on (…)	.24	-.16	.39	Find	.21
10	In which of the sentences below you do not need to cite the source?	.06	.11	.28	Use	.15

Notes. Principal axis factor extraction; Promax rotation with Kaiser normalization; Factor loadings larger than .40 are in bold; RMSEA = .10 [.09, .11]; TLI = .62

* These are aspects that the items were originally generated on the basis of.

$^+$ Extracted

Table A15: Factor correlation matrix

Factor	1	2	3
1		.59	.44
2			.51
3			

3.4 Evaluation of factor solutions from combined samples (pretest sample and graduate sample).

Despite substantially increased sample size, the factorability of the correlations among the IL test items remains poor. Again, a notable feature of the results is that none of the factors in the two- or three-factor solutions seem to align sensibly to the core IL aspects from which test items were derived. And again there seems to be no support for endorsing a single-factor solution, due to the very low proportion of variance explained. Items in the IL test have few and small correlations amongst each other, and this is reflected in generally low factor loadings in all the solutions.

4. Code for the analyses reported above

```
### Exploratory factor analyses of IL-test items ###

library(psych)
library (psychTools)
library(GPArotation)

### Analyses of data from the pilot sample (17 items) ###
pilot_data <- read.file() #Load the file ILpilotdata_forEFA_17items.txt
tet_mat <- tetrachoric(pilot_data) #For visual inspection, scree plot generation and sampling
adequacy assessment
KMO(tet_mat$rho) #Calculate the KMO measure of sampling adequacy
cortest.bartlett(tet_mat$rho, n = 268) #Calculate the Bartlett test of sphericity
scree(tet_mat$rho, factors=FALSE, pc=TRUE) #Generating scree plot from the correlation
matrix

## Factor analysis - One-factor solution ##
fa_pilot_17_1f <- fa(pilot_data, nfactors=1, rotate="promax", scores="regression", fm="pa",
cor="tet") #Calling the factor analysis function with principal axis factoring and promax rotation
with Kaiser normalization.
fa_pilot_17_1f #Display factor analysis output
fa_pilot_17_1f$e.values #Printing eigenvalues

## Factor analysis - Two-factor solution ##
fa_pilot_17_2f <- fa(pilot_data, nfactors=2, rotate="promax", scores="regression", fm="pa",
cor="tet") #Calling the factor analysis function with principal axis factoring and promax rotation
with Kaiser normalization.
fa_pilot_17_2f #Display factor analysis output
fa_pilot_17_2f$e.values #Printing eigenvalues

## Factor analysis - Three-factor solution ##
fa_pilot_17_3f <- fa(pilot_data, nfactors=3, rotate="promax", scores="regression", fm="pa",
cor="tet") #Calling the factor analysis function with principal axis factoring and promax rotation
with Kaiser normalization.
fa_pilot_17_3f #Display factor analysis output
```

```
fa_pilot_17_3f$e.values #Printing eigenvalues

####Analyses of data from the undergraduate (pretest) sample (17 items) ###
ugrad_data <- read.file() #Load the file Undergrad_forEFA_17items.txt
tet_mat <- tetrachoric(ugrad_data) #For visual inspection, scree plot generation and sampling
adequacy assessment
KMO(tet_mat$rho) #Calculate the KMO measure of sampling adequacy
cortest.bartlett(tet_mat$rho, n = 260) #Calculate the Bartlett test of sphericity
scree(tet_mat$rho, factors=FALSE, pc=TRUE) #Generating scree plot from the correlation
matrix

## Factor analysis - One-factor solution ##
fa_ugrad_17_1f <- fa(ugrad_data, nfactors=1, rotate="promax", scores="regression", fm="pa",
cor="tet") #Calling the factor analysis function with principal axis factoring and promax rotation
with Kaiser normalization.
fa_ugrad_17_1f #Display factor analysis output
fa_ugrad_17_1f$e.values #Printing eigenvalues

## Factor analysis - Two-factor solution ##
fa_ugrad_17_2f <- fa(ugrad_data, nfactors=2, rotate="promax", scores="regression", fm="pa",
cor="tet") #Calling the factor analysis function with principal axis factoring and promax rotation
with Kaiser normalization.
fa_ugrad_17_2f #Display factor analysis output
fa_ugrad_17_2f$e.values #Printing eigenvalues

## Factor analysis - Three-factor solution ##
fa_ugrad_17_3f <- fa(ugrad_data, nfactors=3, rotate="promax", scores="regression", fm="pa",
cor="tet") #Calling the factor analysis function with principal axis factoring and promax rotation
with Kaiser normalization.
fa_ugrad_17_3f #Display factor analysis output
fa_ugrad_17_3f$e.values #Printing eigenvalues

####Analyses of data from the merged (undergrad pretest + graduate) sample (21 items) ###
merged_data <- read.file() #Load the file Merged_forEFA_21items.txt
tet_mat <- tetrachoric(merged_data) #For visual inspection, scree plot generation and sampling
adequacy assessment
KMO(tet_mat$rho) #Calculate the KMO measure of sampling adequacy
cortest.bartlett(tet_mat$rho, n = 626) #Calculate the Bartlett test of sphericity
scree(tet_mat$rho, factors=FALSE, pc=TRUE) #Generating scree plot from the correlation
matrix

## Factor analysis - One-factor solution ##
fa_merged_21_1f <- fa(merged_data, nfactors=1, rotate="promax", scores="regression",
fm="pa", cor="tet") #Calling the factor analysis function with principal axis factoring and promax
rotation with Kaiser normalization.
fa_merged_21_1f #Display factor analysis output
fa_merged_21_1f$e.values #Printing eigenvalues

## Factor analysis - Two-factor solution ##
fa_merged_21_2f <- fa(merged_data, nfactors=2, rotate="promax", scores="regression",
fm="pa", cor="tet") #Calling the factor analysis function with principal axis factoring and promax
rotation with Kaiser normalization.
fa_merged_21_2f #Display factor analysis output
fa_merged_21_2f$e.values #Printing eigenvalues
```

```
## Factor analysis - Three-factor solution ##
fa_merged_21_3f <- fa(merged_data, nfactors=3, rotate="promax", scores="regression",
fm="pa", cor="tet") #Calling the factor analysis function with principal axis factoring and promax
rotation with Kaiser normalization.
fa_merged_21_3f #Display factor analysis output
fa_merged_21_3f$e.values #Printing eigenvalues
```

5. References

Aleamoni, L. M. (1976). The relation of sample size to the number of variables in using factor analysis techniques. *Educational and Psychological Measurement, 36*, 879–883. https://doi.org/10.1177/001316447603600410

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates.

Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin, 81*, 358–361. https://doi.org10.1037/h0036316

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. https://doi.org/10.1037/1082-989X.4.3.272

Ford, J. K., Maccallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*, 291–314. https://doi.org/10.1111/j.1744-6570.1986.tb00583.x

Hollis, H. (2018). Information literacy as a measurable construct. *Journal of Information Literacy, 12*(2), 76-88. https://doi.org10.11645/12.2.2409

Nierenberg, E., Låg, T., & Dahl, T. I. (2021). *Replication Data for: Knowing and doing: The development of information literacy measures to assess knowledge and practice* (DataverseNO; Version V2) [Data set, survey and evaluation measures]. https://doi.org/10.18710/L60VDI

Osborne, J. W. (2014). *Best practices in exploratory factor analysis* [Kindle edition]. https://amazon.com

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation, 17(15),* 1-8. https://pareonline.net/getvn.asp?v=17%26n=15

Revelle, W. (2020). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.0.8

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*, 217–222. https://doi.org10.1207/S15327752JPA8003_01

Appendix B

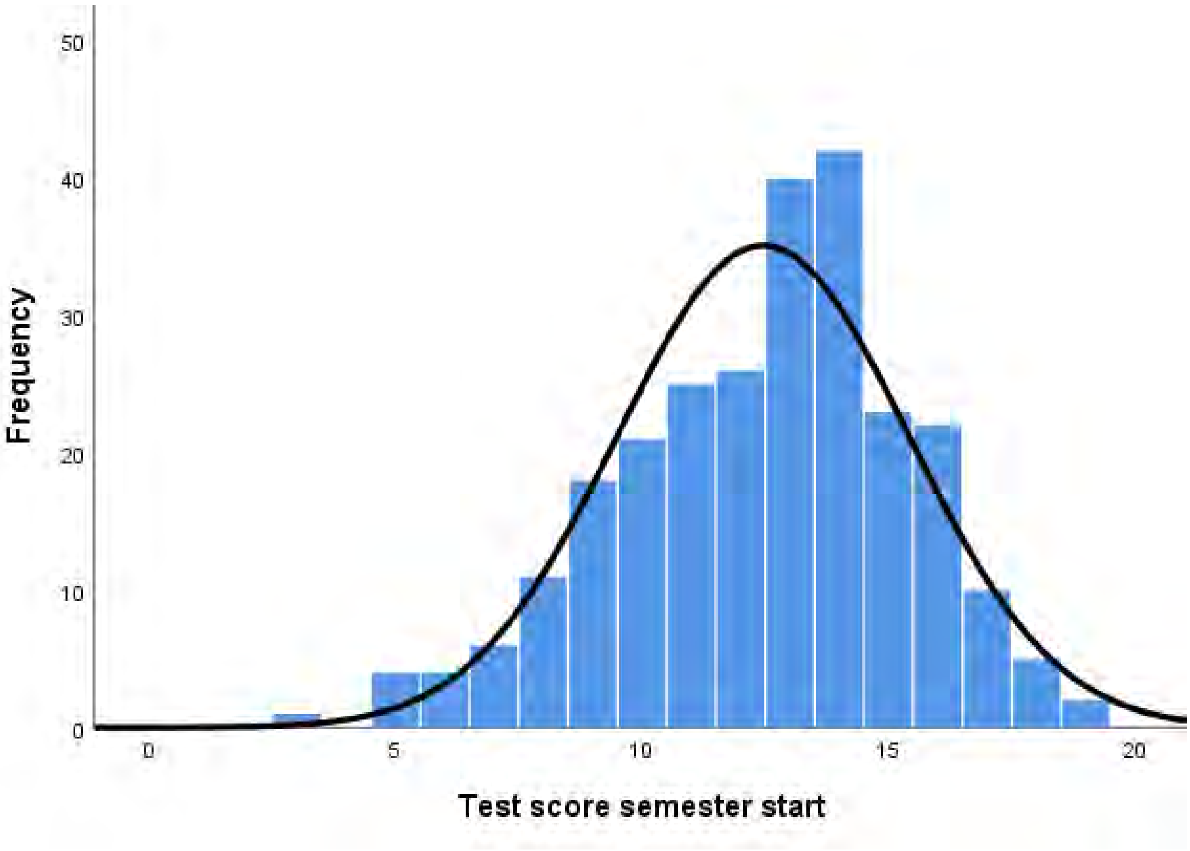

Figure B1: Distribution of undergraduate test scores at semester start

N = 260
Mean = 12.46
Median = 13
Mode = 14
Std. Deviation = 2.953
Skewness = -.424
Std. Error of Skewness = .151
Kurtosis = -.034
Std. Error of Kurtosis = .301

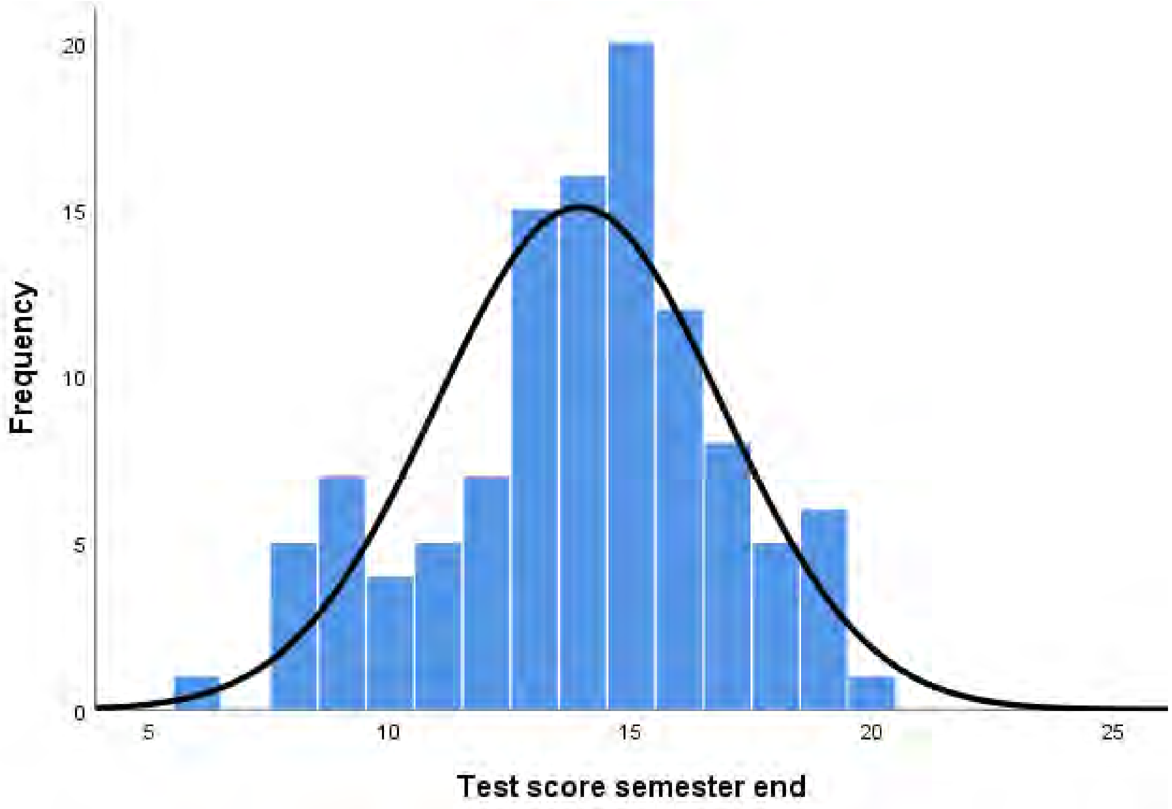

Figure B2: Distribution of undergraduate test scores at semester end

N = 112
Mean = 13.92
Median = 14
Mode = 15
Std. Deviation = 2.966
Skewness = -.378
Std. Error of Skewness = .228
Kurtosis = -.217
Std. Error of Kurtosis = .453

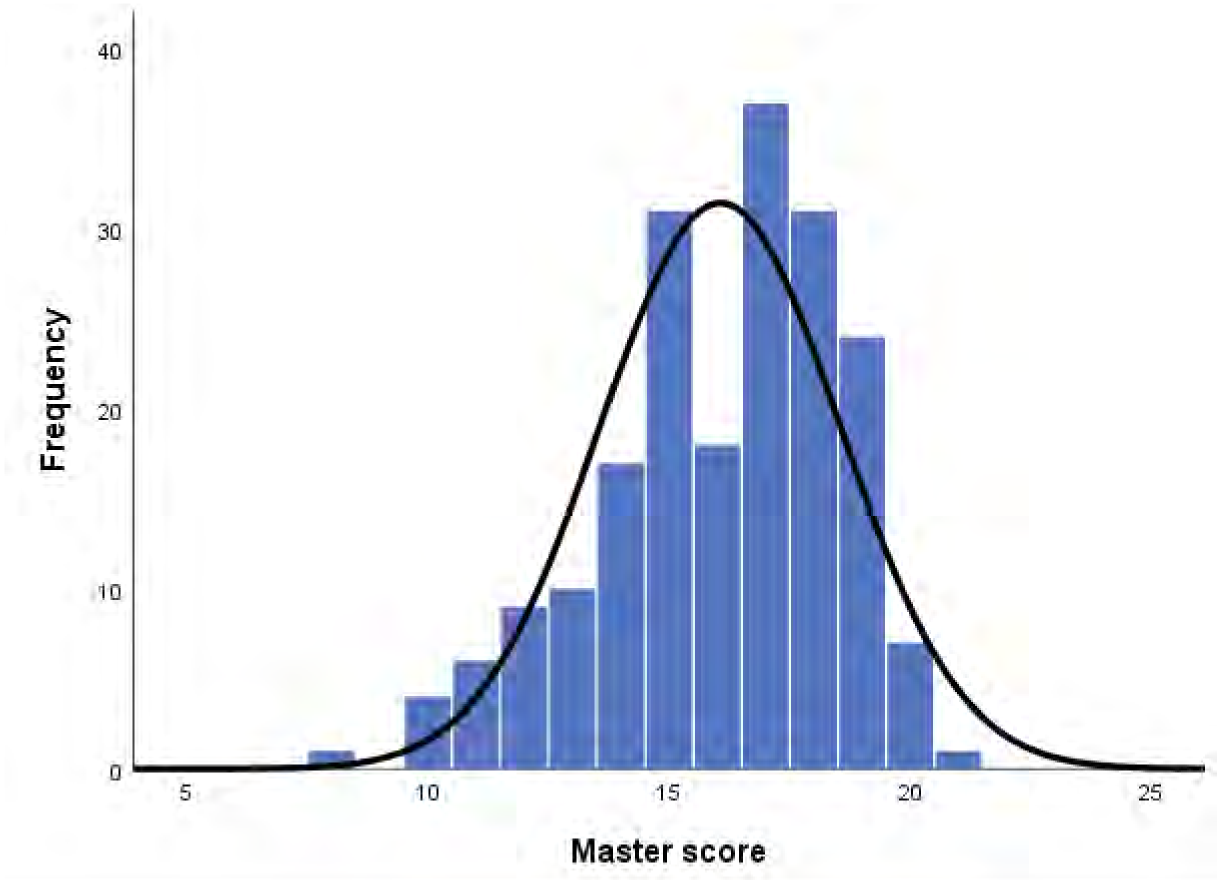

Figure B3: Distribution of Master's student scores

N = 196
Mean = 16.06
Median = 17
Mode = 17
Std. Deviation = 2.485
Skewness = -.620
Std. Error of Skewness = .174
Kurtosis = -.034
Std. Error of Kurtosis = .346

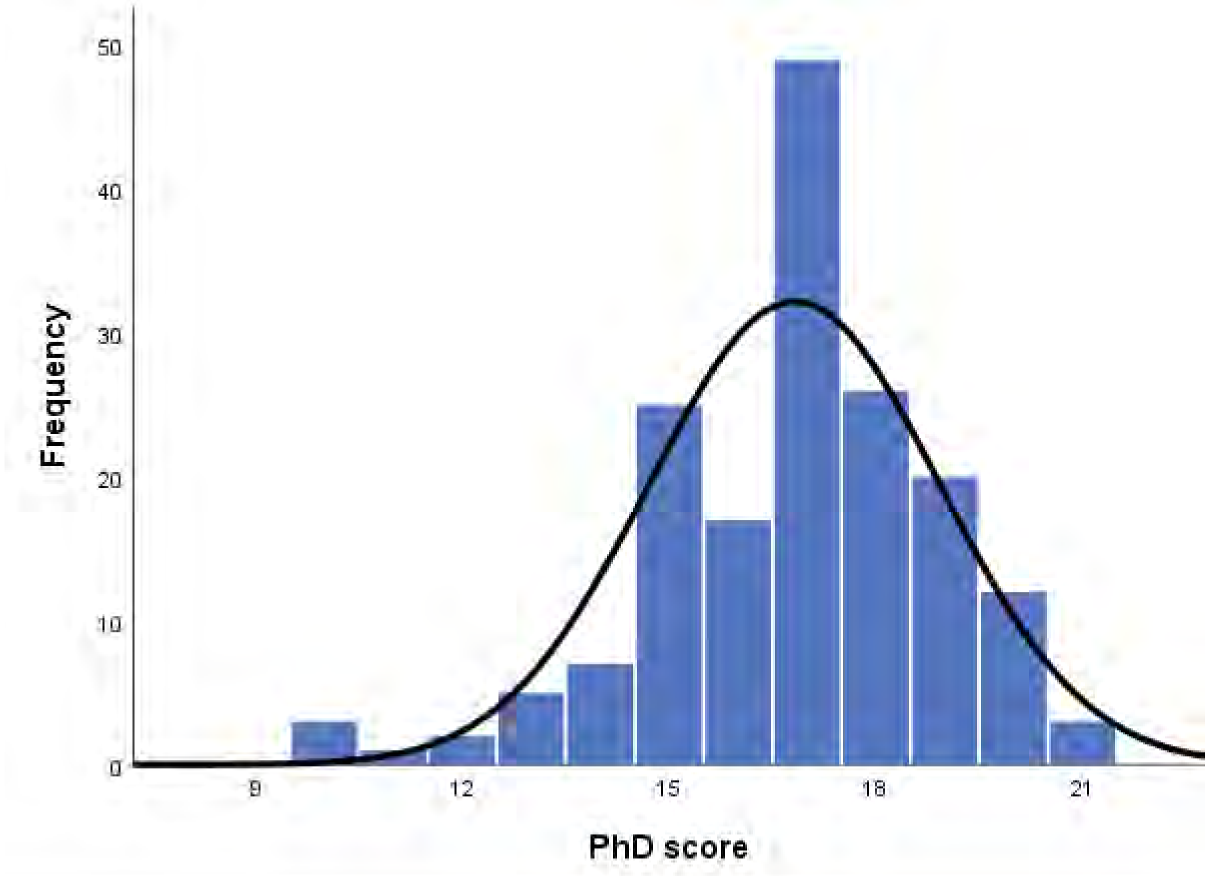

Figure B4: Distribution of PhD student scores

N = 170
Mean = 16.82
Median = 17
Mode = 17
Std. Deviation = 2.106
Skewness = -.711
Std. Error of Skewness = .186
Kurtosis = 1.044
Std. Error of Kurtosis = .370

Appendix C

1. Correlations

Correlations for seven study variables are reported in Table C1:

- KNOW
    - Information literacy knowledge test, semester-start score (KnowStart)

- DO
    - Source Evaluation total score (SourceEval Total Score)
    - Source Evaluation quality score (SourceEval Quality)
    - Source Evaluation variety score (SourceEval Variety)
    - Source Evaluation frequency score (SourceEval Frequency)
    - Source Use score (SourceUse Score)

- EXPERIENCE
    - Number of semesters of previous higher education (Semesters HE)

Table C1: Correlations for study variables, in Pearson's r

Variable	*n*	1	2	3	4	5	6	7
1. KnowStart	260	–						
2. SourceEval Total Score	93	.195	–					
3. SourceEval Quality	93	.272**	.466**	–				
4. SourceEval Variety	93	.150	.932**	.359**	–			
5. SourceEval Frequency	93	.184	.940**	.300**	.943**	–		
6. SourceUse Score	87	.308**	.197	.157	.141	.166	–	
7. Semesters HE	260	.150*	-.033	.022	-.032	-.052	-.105	–

Notes:
$^{*}p < .05$
$^{**}p < .01$

1.1 Comments on correlations

Aside from correlations between the different source evaluation scores, there are three significant correlations between the analyzed variables:

- The source evaluation quality score and the source use score are positively correlated with the Information literacy (IL) knowledge at semester-start scores. These correlations are statistically significantly, but not strong.
- There is a positive correlation between the Number of semesters of previous higher education and IL knowledge at semester-start scores. This correlation is statistically significant, but weak.

2. Regressions

In order to follow up on the associations reported in the correlation matrix, and specifically, to attempt to control for higher education experience, we performed two multiple regressions using IL knowledge test scores and number of semesters of higher education as predictors, and the source evaluation quality score and source use scores as dependent variables.

For both of these analyses there were no indications of multicollinearity or influential cases, errors are independent, and there was no obvious heteroscedasticity. Residuals showed some deviation from normality, particularly for the analysis reported in Table C2.

Table C2: Regression coefficients of IL knowledge test scores on Source Evaluation quality score

Variable	Model 1			Model 2		
	B	***β***	***SE***	***B***	***β***	***SE***
Constant	1.71**		.34	1.71**		.34
IL knowledge semester start score	.07**	.27	.03	.07**	.27	.03
Numbers of semesters in HE				.00	.00	.08
R^2	.07			.07		
ΔR^2				.00		

Note. *N* = 93. In Model 1, we entered the IL test scores. In Model 2, we entered Number of semesters of higher education as a control variable.
***p* < .01.

Table C3: Regression coefficients of IL knowledge test scores on Source use score

Variable	Model 1			Model 2		
	B	***β***	***SE***	***B***	***β***	***SE***
Constant	1.07**		.56	1.73**		.55
IL knowledge semester start score	.13**	.31	.04	.13**	.32	.04
Numbers of semesters in HE				-.17		.14
R^2	.10			.11		
ΔR^2	.02					

Note. *N* = 87. In Model 1, we entered the IL test scores. In Model 2, we entered Number of semesters of higher education as a control variable.
***p* < .01.

2.1 Comments on regressions

None of these models account for much of the variance in the dependent variables (scores on do-measures). Nevertheless, as an illustrative example, and assuming the models generalize, a standardized coefficient of just above .3 between knowledge test at semester start scores and source use scores, allows us to predict almost a third of a standard deviation increase in the ability to use sources for every increase of one standard deviation in IL knowledge. In other words, an increase of just under three points on the knowledge test translates to a 0.34 point increase on the source use score (which has a maximum score of five). While not impressive, it is not entirely trivial.

Note that the attempted control for number of semesters is weak, due to the restricted range of this variable in this sample.