Measuring psychology students' information-seeking skills in a situational judgment test format: Construction and validation of the PIKE-P Test

Tom Rosman and Anne-Kathrin Mayer

ZPID, Leibniz Institute for Psychology Information and Documentation, Trier

Günter Krampen

University of Trier and ZPID, Leibniz Institute for Psychology Information and Documentation, Trier

Recommended citation:

Rosman, T., Mayer, A.-K., & Krampen, G. (2016). Measuring psychology students' information-seeking skills in a situational judgment test format: Construction and validation of the PIKE-P Test. *European Journal of Psychological Assessment, 32,* 220-229. doi: 10.1027/1015-5759/a000239.

Correspondence concerning this article should be addressed to Tom Rosman, ZPID, Leibniz Institute for Psychology Information, 54296 Trier, Germany.

E-mail: rosman@zpid.de

Summary

Three studies were conducted to develop a test for academic information-seeking skills in psychology students that measures both procedural and declarative aspects of the concept. A skill decomposition breaking down information-seeking into ten sub skills was used to create a situational judgment test with 22 items. A scoring key was developed based on expert ratings ($N = 14$). Subsequently, the test was administered to two samples of $N = 78$ and $N = 81$ psychology students. Within the first sample, the scale reached an internal consistency (Cronbach's Alpha) of $\alpha = .75$. Scale validity was investigated with data from the second sample. High correlations between the scale and two different information search tasks ($r = .42$ to $.64$; $p < .001$) as well as a declarative information literacy test ($r = .51$; $p < .001$) were found. The findings are discussed with regard to their implications for research and practice.

Keywords: Information-seeking, Information Literacy, Situational Judgment Test, Procedural knowledge, Declarative knowledge

1 Introduction

Information Literacy is defined as a set of abilities necessary to recognize an information need and to subsequently locate, evaluate, and use the needed information (Association of College and Research Libraries [ACRL], 2000). Even though information evaluation and use should not be disregarded, the ability to develop and implement information-seeking strategies is a prerequisite for all subsequent information processing (Goldhammer, Kröhne, Keßel, Senkbeil, & Ihme, 2014). Moreover, universities are the places where many students get in touch with academic information-seeking for their first time. Without any prior knowledge, they have to acquire skills to master a multitude of complex information retrieval methods. As such information-seeking skills closely relate to student achievement (e.g., Bowles-Terry, 2012), it seems rather worrying that their assessment is still in its infancy (Walsh, 2009). The few existing standardized instruments are multiple-choice tests that “cannot assess the effectiveness of student search skills in real life situations” (Dunn, 2002, p. 27). Measuring information-seeking strategies through more ecologically valid information search tasks, on the other hand, is time consuming and exhibits statistical inconveniences. The present article introduces a test that tries to overcome these difficulties by drawing on a situational judgment format and claims to assess “procedural” as well as “declarative” aspects of knowledge about information-seeking.

In educational psychology, declarative knowledge is defined as knowledge about certain facts and events, thus representing factual knowledge (“knowing what”). It can be achieved through explicit learning, for example by memorizing facts. In contrast, procedural knowledge represents context-specific knowledge about how to behave in certain situations or about how to solve certain problems (“knowing how”). This form of knowledge also includes unconscious and tacit components (Stemler & Sternberg 2006). As the information literacy standards of the Association of College and Research Libraries (2010) include both more declarative (“Distinguishes between empirical study and literature review”, p. 2) and more procedural (“Retrieves scholarly journals, books, and sources appropriate to the inquiry”, p. 4) components, one can assume that information-seeking requires both forms of knowledge: A student searching for academic information needs, for example, declarative knowledge on the existence and scope of relevant search engines, and procedural knowledge on how to operate them.

Apart from self-report measures that have been widely criticized for producing biased test scores, three different approaches to the assessment of information-seeking skills can be identified. First, a few established *multiple-choice achievement tests* for information literacy exist (e.g., Information Literacy Test [ILT]: Wise, Cameron, Yang, Davis, & Russell, 2009; Leichner, Peter, Mayer, & Krampen, 2013). These are mostly broad and generic inventories claiming to measure several ACRL-standards with different subscales. Their rather short items seem to primarily assess declarative knowledge. For example, while working on the subscale *Access* (second ACRL-standard) of the ILT, subjects are asked how the database which includes books from a specific library (the library catalog) is called. The main advantage of these tests is their easy administration and scoring. Moreover, they effectively prohibit faking. However, isolating the respective subscale (e.g., subscale *Access* for the ILT) to measure information-seeking skills constitutes a risky endeavor, especially since the factor structure of these tests has not been confirmed empirically. Additionally, even though reliability indices are sometimes reported, convergent validity has—to the best of our knowledge—never been tested for any of these instruments.

A second test format constitutes so-called *information search tasks* (e.g., Ivanitskaya, O’Boyle, & Casey, 2006; Leichner, Peter, Mayer, & Krampen, 2014): In a real-world setting (i.e., on a computer with access to bibliographic databases), subjects are instructed to find a scientific article about a certain subject, and their search results are evaluated. Even though some declarative knowledge is necessary to fulfill these tasks (e.g., about the scope of certain databases), they mainly measure procedural knowledge, and thus consitute the most ecologically valid indicator of information-seeking skills. Unfortunately, these tests have limitations, too: First, conducting extensive search tasks is time-consuming (Dunn, 2002). Answering one single item in the test by Leichner et al. (2013) takes up to ten minutes, which diminishes the reasonable number of test items and thus reduces test reliability. Scoring is time-consuming as well, and objectivity is difficult to achieve because the search results of each subject have to be assessed individually to evaluate whether subjects found an article that matches all criteria given in the instruction. Finally, test results are influenced by contextual factors such as the availability of certain bibliographic databases in the specific test setting.

A third method of assessing information-seeking skills are *simulation tasks* (e.g., Goldhammer et al., 2014). Such tasks are based on the re-creation of relevant database features in a simulated environment, attenuating many of the search tasks' limitations: They can be scored automatically and allow a sophisticated analysis of the search process as such, for example through the inspection of reaction times. Additionally, practically identical simulations can be administered via USB stick on almost every computer (Greiff, Wüstenberg, Holt, Goldhammer, & Funke, 2013), which reduces the influence of contextual factors. However, even though simulations are very well suited to investigate procedural knowledge on one specific (simulated) database interface, they may not account for the whole spectrum of available search tools, and may not capture the whole complexity of information-seeking processes (e.g., they neglect planning behavior). Additionally, their development is time-consuming, especially when re-creating complex environments like bibliographic databases. Finally, no simulation tasks for academic information-seeking in higher education have yet been constructed.

To sum up, all of the presented assessment methods have their raisons d'être. On the downside, however, they either do not measure procedural knowledge (multiple-choice tests), exhibit scoring problems (information search tasks), or do not account for the whole breadth and depth of information search processes (simulation tasks). A promising alternative approach to the measurement of both procedural and declarative knowledge consists in formulating items in a *situational judgment test* format (SJT; e.g., Motowidlo, Hooper, & Jackson, 2006). SJT items consist of a description of a situation followed by questions on the behavior of subjects in that specific situation. They are usually presented in a multiple-choice format, which makes them easy to administer and score. However, compared to information search tasks, (single) SJT items are not suited to assess the whole depth of information-seeking processes, because multiple-choice response alternatives cannot become indefinitely complex, and because test takers cannot interact with the test material (e.g., "change" the presented situation through their actions; Goldhammer et al., 2014). Furthermore, distractors have to be chosen very carefully in order to minimize guessing. Finally, construction of these tests requires a more elaborate theoretical framework than the design of information search tasks: Prior to scenario development, key situations in the information-seeking process have to be identified, followed by the extraction of specific approaches to handle these situations.

The ability to locate and access scientific information is often characterized as a set of several sub skills, and researchers have begun to decompose it accordingly. As stated earlier, the well-known ACRL-standards refer to different sub skills required for information-seeking in the domain of psychology (ACRL, 2010). With the purpose of identifying, describing, and classifying the (sub) skills necessary to conduct more complex information searches, Brand-Gruwel, Wopereis, and Vermetten introduced the term *skill decomposition* in 2005. As their empirically derived skill decomposition however does not target psychology students, the authors of the present paper screened both approaches for similarities in order to develop a coherent *skill decomposition* of information-seeking in psychology. To ensure content validity of the test, ten sub skills covering all relevant areas of academic information-seeking, were carefully selected (see Table 1). These can be classified into the broader categories *Development of Search Strategies* and *Application of Search Strategies*.

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

With this skill decomposition as a theoretical basis, we developed the PIKE-P test (Procedural Information-seeking Knowledge Evaluation—Psychology version) designed to measure information-seeking skills in psychology students by drawing on an SJT format which allows the measurement of both procedural and declarative knowledge and circumvents the problems associated with multiple-choice tests and information search tasks.

2 Materials and Methods

Scale development was subdivided in four phases: Item development, scoring key development, empirical evaluation of the items, and validation of the final scale.

Item development according to our skill decomposition

During the first phase, a group of three researchers created an initial pool of 40 items. Items were carefully chosen to represent our skill decomposition, and their wording was modified several times until all three researchers agreed over their appropriateness. In accordance with the SJT method depicted above, each item began with the description of a key situation in the information-seeking process (the so-called scenario), and was followed by four approaches to the situation that could be rated on a 5-point Likert-Scale ranging from *not useful at all* to *very useful*. For sub skills that primarily focused on procedural knowledge (e.g., utilization of online thesaurus), scenarios were followed by four rather long and complex descriptions of possible approaches (see Table 2 for a sample item). Such genuine SJT-items have been demonstrated to be sensitive to procedural and even tacit knowledge (e.g., because they describe multiple steps of the search process; Stemler & Sternberg 2006; Goldhammer et al., 2014). On the other hand, items dealing with sub skills that primarily focus on declarative knowledge (particularly the sub skills on source or publication type selection) included four short, often single-worded approaches. Even though such items could also be reworded into a simpler multiple-choice format (see Tables 3 and 4), we decided to retain their SJT format because working on real-life scenarios constitutes a motivating factor. For both types of items, the four situational approaches differed in their appropriateness to deal with the presented situation: Some could be considered as very useful, others as moderately useful, whereas some approaches were not useful at all.

----- Insert Tables 2, 3, and 4 here -----

Expert study to gain scoring rules

To gain an appropriate scoring key, a preliminary 40-item version of the test was administered online to N = 14 subject matter experts. The sample consisted of n = 7 reference librarians working in university libraries throughout Germany, and n = 7 psychology lecturers working at the University of Trier. Gender was distributed evenly across the sample. Mean age was M = 42.85 (SD = 13.33) years, and subjects reported an average information search experience of M = 15.79 (SD = 10.12) years. Throughout all three studies, the scale was administered in German language. The experts were asked to rate (on the five point scale; see above) how appropriate the approaches for dealing with the situational purposes of the 40 items were. They could also make remarks on the items in a commentary field. Based on these remarks, a few minor corrections to the wording of the items were made.

The scoring key was developed based on a method by Artelt, Beinicke, Schlagmüller, and Schneider (2009), which focuses on the comparison of pairs of responses to two different approaches (i.e., how to deal with the purpose in the scenario) rather than on absolute values on the rating scale. With four approaches (A to D), a maximum of six pairwise comparisons can be conducted per item (AB, AC, AD, BC, BD, and CD). For each pair of approaches, it is considered whether the response on the rating scale to one approach is higher (or lower) than the response on the rating scale to the other approach: If a subject prefers approach A over B, his or her score is increased by one point if (and only if) the experts consistently preferred approach A over B, too. Thus, the preferences of the experts provide the scoring key for the scoring of the subjects. For example, referring to the sample item presented in Table 3, it was found that 93 % of the experts preferred B (searching PsycINFO) over A (searching the library catalog). Based on that information, the pairwise comparison AB was added to the scoring key: Only if subjects give a higher rating on the five-point rating scale for approach B (e.g., 4) than for approach A (e.g., 2), their score is increased by one point. If both approaches are rated equally (e.g., 4), no point is awarded (for some pairwise comparisons, this was liberalized later; see paragraph on the empirical evaluation of the items). An advantage of this method is that the impact of response biases (like extreme bias or modesty bias) is reduced. For example, if the scoring based on absolute values (+5 when the correct approach is rated with 5), subjects with a general tendency for extreme ratings would score higher than subjects who favor the middle categories of the scale. This problem does not occur with pairwise comparison scoring. Additionally, pairwise scoring has the advantage that items stay valid if new search tools or functions are developed. Even though means of the responses to the approaches would diminish (which would cause serious issues with a scoring key based on absolute values), the pattern of ratings would stay identical and no changes to the scoring key would have to be made.

As agreement between raters varied over different pairwise comparisons, a cut-off criterion of 70 % was set: If, for a specific pairwise comparison, more than 70 % of experts agreed that one approach is better suited than the other, this pairwise comparison was added to the scoring key. The rationale behind the choice of this rather liberal criterion is as follows: Previous research has shown that information problems can be resolved in different ways and that novices might employ less efficient (but still effective) approaches to information-seeking than experts. For example, experts have been shown to invest more time in planning their search (Brand-Gruwel et al., 2005), use more sophisticated search strategies (Macedo-Rouet, Rouet, Ros, & Vibert, 2012), and generate longer and more complex search queries (Sutcliffe, Ennis, & Watkinson, 2000). Nevertheless, novices might also yield satisfactory results through the use of simpler approaches; for example if they invest more time in search query iteration (Sutcliffe et al., 2000). To account for this, the situational approaches of the PIKE-P do not simply reflect “right” and “wrong” approaches, but also include moderately useful (“simpler”) approaches. This however entails more room for subjectivity and reduces expert agreement. For example, experts applying a rather strict criterion might rate moderately useful approaches as just as useless as the truly bad approaches, while other experts are more liberal and accept them as an alternative to “optimal” approaches. As expert agreement was not consistent enough to meet their more conservative 80 % criterion, Artelt et al. (2009) eventually had to drop their middle categories and switch to a dichotomous (only right and wrong strategies) test format. Due to the significance of such middle categories for the assessment of information skills (information problems can be resolved in multiple ways that nevertheless differ in their efficiency), we chose to retain them and lower the cut-off criterion instead.

Applying these considerations, 12 of the initial 40 items were eliminated from the test because the experts differed too much in their preferences. The preliminary scoring key of the remaining 28-item version comprised 3 to 6 pairwise comparisons per item and 112 pairwise comparisons in total.

Empirical evaluation of the items

The third phase of scale development was carried out to evaluate the test in a real-world setting, select the most adequate items, and eliminate pairwise comparisons with low discriminatory power. With this purpose, the 28-item-scale was administered to a pilot sample of N = 78 psychology students at the University of Trier. Data collection was conducted in groups of 4 to 28 students who were paid for their participation. About three quarters (76 %) of the subjects were female, a distribution that is typical for German psychology students (Wentura, Ziegler, Scheuer, Bölte, Rammsayer, & Salewski, 2013). Mean age was M = 23.87 (SD = 3.32) years. Two third (66 %) were undergraduates and one third (34 %) were seeking a master's or equivalent degree.

After data collection, the preliminary scoring key was applied to the data to gain initial test scores. Thereafter, part-whole-corrected correlations between single pairwise comparisons and the total test score were calculated and inspected graphically through scatterplots in SPSS^{TM} Version 20 (abscissa: single pairwise comparisons; ordinate: item-corrected test score). To obtain a scoring key that differentiates well between subjects of the test's target group (psychology students of all semesters), this information was then used to eliminate pairwise comparisons with low discriminatory power (i.e. low correlations with the total test score). Moreover, we slightly liberalized awarding of score for a few pairwise comparisons: If one approach was just *slightly* more useful than the other (as could be seen in the scatterplots as well as in rather inconsistent expert ratings), subjects score also increases when they rate both approaches as equally useful. We thereby took great care not to change rank orders of approaches: Approaches rated as “wrong” by the experts would under no circumstances become “right” in the scoring key and vice-versa. Pearson correlations between test scores generated by the preliminary and the final scoring keys support the assumption that these minor changes did not affect the test scores substantially: In the expert sample, both scores correlated by $r = .88$ ($p < .001$); in the pilot sample, the correlation was even higher with $r = .95$ ($p < .001$).

Finally, six items with a low discriminatory power were eliminated, reducing test duration to 25 minutes. It was thereby taken care that the final test includes all ten sub skills (two to three items per sub skill) to ensure content validity. The final test consists of 22 items, with a final scoring key comprising 3 to 5 pairwise comparisons per item and a maximum score of 86 points[1].

Validation study

In order to establish construct validity, a third study was conducted with psychology students at Saarland University in Saarbrücken. Data collection was conducted in groups of 8 to 15 students. Participants were again paid for their participation and sample size was (after elimination of one subject due to technical problems) $N = 81$. Gender was distributed very unevenly with 91 % female and 9 % male participants. Mean age was $M = 22.33$ ($SD = 2.99$) years. Two third (67 %) of subjects were undergraduates and one third (33 %) were in a master’s degree program.

[1] The full test as well as an SPSS syntax file computing test scores is available on request from the first author.

In order to judge the convergent validity of the final PIKE-P test, participants also completed the ILT-P, a multiple-choice test that assesses declarative information literacy knowledge (Leichner et al., 2013), and two different types of information search tasks as validation criteria. The first task type targeted full text acquisition: Subjects were given nine APA-Citations (books, journal articles, etc.) and had to locate the corresponding full text. Some publications were available online, some were only available in the local library, whereas others required the use of interlibrary loan. Subjects were asked to either indicate an URL or a library signature, or mention that interlibrary loan was required. A time limit of 30 minutes was set to solve the nine items. For each participant, we scored how many relevant publications he/she had found. The second type of tasks was based on the work by Leichner et al. (2014). Subjects worked on three information search tasks of increasing difficulty that were chosen randomly from a pool of three tasks per difficulty level (e.g., medium level: "Find two longitudinal studies of risk factors for generalized anxiety disorder published after 2005"). Again, time to solve each problem was limited to 4 to 10 minutes, depending on task difficulty. Upon completion, subjects were required to document both title and (first) author of two corresponding publications. Furthermore, they had to answer questions regarding their course of action during the information search (e.g., "Which search tool did you use?"). To assess whether the documented publications fitted task requirements, they were rated based on a standardized scoring rubric reflecting outcome quality (*outcome score*). The information derived from the additional questions was used to calculate a *process score* indicating the quality of the search process. In addition, three self-report questions about prior information search experience had to be answered (e.g., "How often do you conduct literature searches about a specific subject?"; 5-point Likert-Scale response format).

3 Results

In order to assess the psychometric properties of the scale, the final scoring key was applied to all three datasets (expert, pilot, and validation sample). Table 5 summarizes descriptive statistics of the scale for all three samples.

----- Insert Table 5 here -----

As expected, experts scored considerably higher than students, with nearly all individual scores in the top quarter of the distribution. In the pilot study, the scale yielded an internal consistency (Cronbach's Alpha) of α = .75 and a Spearman-Brown split-half-reliability coefficient (odd-even method) of r_{xy} = .78, with item discrimination power ranging from r_{itc} = .02 to .55. In the validation study, reliability coefficients were slightly lower, with an internal consistency of α = .72 and a Spearman-Brown split-half-reliability of r_{xy} = .75. For this sample, item discrimination power ranged from r_{itc} = .05 to .45.

Additional tests were performed to investigate scale validity. The pilot sample comprised *N* = 21 subjects that had taken part in an information literacy course two weeks before data collection. This course used a blended-learning approach to foster students' skills in finding and evaluating psychology-related information, including training in the practical use of bibliographic databases. As all course participants were freshmen and sophomores, they were compared to the *N* = 19 freshmen and sophomores in the pilot sample who had not undergone training. Trained participants achieved much higher PIKE-P scores (*M* = 60.81; *SD* = 7.87) than non-trained (*M* = 46.53; *SD* = 9.20) participants ($t_{(38)}$ = 5.29; *p* < .001). This finding can be interpreted as a first validity indicator of the PIKE-P test, because the intervention (the training) led to higher scores.

To further investigate construct validity, data from the validation study were taken into account. Table 6 shows descriptive statistics and intercorrelations of all study variables.

----- Insert Table 6 here -----

The results of the validation study reveal high convergent validity of the PIKE-P test. Both outcome and process score of the information search tasks highly correlate with PIKE-P scores. For the full text acquisition tasks, correlations are slightly lower, which might be due to full text acquisition being only marginally represented (by one single sub skill) in the PIKE-P test. As for relations with a primarily declarative measure of information literacy—the ILT-P—correlations again indicate a moderate to high convergent validity. Finally, positive correlations with study progress show that more advanced (and therefore experienced) students produce higher test scores. To further investigate convergent validity, a principal component analysis was conducted with the mean scores of PIKE-P, ILT-P, full text acquisition tasks, and both the outcome and the process score of the information search tasks (see Table 6). All tests loaded on one factor (eigenvalue = 3.09; loadings between .65 and .85), which explained 62 % of total variance. PIKE-P (*a* = .83), search task outcome (*a* = .82) and process score (*a* = .85) loaded high on this factor, while ILT-P (*a* = .76) and full text acquisition tasks (*a* = .65) had lower yet substantial loadings.

Further calculations were performed to assess incremental validity of PIKE-P over alternative information literacy indicators. As performance in information search tasks can be seen as the most ecologically valid indicator for information-seeking skills, we conducted regressions of the two search task scores on PIKE-P scores while controlling for ILT-P scores, self-reported information search experience, and study progress. With all control variables in the equation, PIKE-P scores remained significant predictors for both outcome and process scores of the search tasks (see Table 7). This shows that the PIKE-P test indeed explains incremental variance in information search task scores over ILT-P scores, self-reported search experience, and study progress.

----- Insert Table 7 here -----

4 Discussion

The purpose of this article was to introduce a new test (PIKE-P; Procedural Information-seeking Knowledge Evaluation—Psychology version) to assess information-seeking skills in psychology students. To ensure content validity, the authors conducted a skill decomposition breaking down information-seeking into ten sub skills. Test items are presented in a situational judgment format, comprising a scenario-part and four approaches which differ in their appropriateness to resolve the presented situation. During the multi-phase construction process, 40 items were first administered to a sample of *N* = 14 subject matter experts. With that data, an initial scoring key was developed, which focuses on the comparison of pairs of responses to the different situational approaches: For each pair, it is considered whether subjects prefer the approach that at least 70 % of the experts preferred; if this is the case a scoring point is given. Items with low expert agreement were eliminated. In a second study, the remaining items were administered to *N* = 78 psychology students to select the most adequate items and refine the scoring key. A third study dealt with scale validation. The final scale comprises 22 items (2-3 items per sub skill).

Overall scale reliability (internal consistency and split-half-reliability) is satisfactory in both student samples (α = .72 - .75; r_{xy} = .75 - .78). As information literacy is a multi-faceted concept (Andretta, 2005), this somewhat lower reliability is likely due to information-seeking skills varying intra- and inter-individually for the different sub skills. In that sense, some students should know how to formulate adequate search queries, others how to use limiters or online thesaurus, and only few should have comprehensive information search knowledge. Impaired reliability is the obvious consequence. Measuring all sub skills with five or more items each would remedy this situation, as one could assess subscale reliability of a multidimensional instrument. However, concerned about the detrimental effects of excessive test length on motivation and compliance, we decided to construct a short, economical scale with the disadvantage of having to accept slightly lower reliability levels.

For scale validity, empirical evidence is convincing: Correlations indicating convergent validity were found between test scores and both outcome and process scores of the information search tasks. As such tasks can be seen as the most “natural” (or ecologically valid) indicator for information-seeking skills (because successful information searches require high information-seeking skills per se), they are best suited to validate the PIKE-P test. Consequently, these high correlations make a strong point for the test’s construct validity: Subjects with high test scores indeed have a higher chance of finding publications on a given topic (outcome scores) and use more efficient and effective search strategies (process scores). For the full text acquisition tasks, correlations are slightly lower, but nevertheless indicate that students with higher PIKE-P scores have a higher chance of finding full texts of given publications. To further strengthen convergent validity, another information literacy test (the ILT-P) was administered. Although this test primarily focuses declarative information-seeking knowledge and even includes items on information evaluation, the correlation between both tests indicates moderate to high criterion validity. After introducing ILT-P scores, self-reported information search experience, and study progress as control variables, all correlations remained significant, which indicates incremental validity of the PIKE-P over these measures. Furthermore, significant differences between information literacy training participants and non-participants demonstrate sensitivity to changes in information literacy levels. Finally, a principal component analysis of all tests included in the validation study indicated a single factor solution. Even though we initially assumed that some of these tests would primarily measure procedural knowledge whereas others would relate more strongly to declarative aspects of information-seeking, this was not shown by the principal component analysis. This suggests that a strict differentiation between declarative and procedural knowledge might not be possible in the assessment of information-seeking skills and that both knowledge components play a role in information-seeking. Nevertheless, the presence of a one-factor solution indicates that all tests included in the validation study principally measure the same concept. Moreover, PIKE-P and research tasks having the highest loadings on this factor can be seen as yet another validity indicator of the PIKE-P.

The results of our validation study are particularly important as, though very desirable, empirical testing of scale validity is not common at all in information literacy research. For the PIKE-P, however, convergent validity has been thoroughly tested. Moreover, our results show that the PIKE-P test is indeed well-suited to replace information search tasks as a shorter and more economic measure for information-seeking skills. Additionally, the test does not require specific library infrastructure (e.g., access do bibliographic databases), and could even—at least in theory—be administered in a paper-pencil format. Another strength of the PIKE-P test is that all items are, due to their structure and formulation, largely independent from situational conditions such as the availability of specific database interfaces. For example, subjects are not asked about specific features of specific databases, but about global functions that nearly all these databases possess (e.g., limiters or online thesaurus). The student studies having been conducted at two different universities provide evidence for this independence from situational conditions. In fact, both university libraries offer different information infrastructures: To access the bibliographic database PsycINFO™, the University of Trier uses the search interface of OvidSP™, whereas Saarland University uses EBSCOhost™. Despite these differences, psychometric indicators, means, and standard deviations of the test do not differ substantially throughout both studies. This finding is important in light of the short lifecycles of today's information infrastructure, as the items will not become obsolete if the database interface is modified. Additionally, as the test has shown to be reliable with both EBSCOhost™ and OvidSP™ (which are among the biggest PsycINFO™ interfaces), there is a good case to believe that the test also proves reliable internationally.

Nevertheless, the PIKE-P also has some limitations. First, we have to acknowledge that it only measures information-seeking skills, neglecting information evaluation and use as well as ethical aspects. Because of the strong focus on information-seeking by most library interventions, this should not constrain the application of the test in such contexts (Kompetenznetzwerk für Bibliotheken des Deutschen Bibliotheksverbands, 2013). Regarding educational research or more elaborate interventions, it should however be noted that the test is only sensitive to certain aspects of information literacy. Especially the importance of information evaluation should not be neglected, and constructing an instrument that measures both declarative and procedural knowledge about information evaluation and use is a major challenge that future research needs to tackle.

Second, the pairwise scoring might entail artificial dependencies because some situational approaches (in particular, very good and very bad approaches), are more heavily weighted (through an increased number of corresponding pairwise comparisons in the scoring key) than moderately useful approaches. As “mistakes” in approaches which exhibit such a clear differentiation between right and wrong indicate rather low literature search competencies, this uneven weighting of approaches is however justified.

A third limitation concerns generalizability: We have to point out that in the validation study, gender was even more strongly skewed (only 9 % males) than is usually the case in German psychology students (Wentura et al., 2013). This might impair the generalizability of our findings to a more general population of psychology students.

A fourth limitation might be that some items—particularly the two items of the fourth sub skill—partly measure general psychological knowledge and not just knowledge related to information-seeking. This implies that advanced students might score higher than novices solely due to better terminology knowledge and orientation to the subject. As the rewording of search terms and the activation of prior knowledge plays a very central role in information-seeking (Brand-Gruwel et al., 2005; ACRL, 2010; see also Table 1), we chose to retain such items. To limit the impact of the issue, we constructed items that only required basic psychological vocabulary susceptible to be mastered by almost any psychology student (e.g., knowledge about differences between classical and operant conditioning; see Table 4). The rather large differences in PIKE-P scores between information literacy training participants and non-participants—all of the same cohorts and thus with similar levels of psychological vocabulary—provide evidence for the test’s sensitivity to information-seeking knowledge.

To sum up, the PIKE-P constitutes (up to now) the most comprehensive measure for information-seeking skills in psychology students, investigating both procedural and declarative knowledge on information-seeking in an ecologically valid and psychometrically sound way. Its simple administration and scoring allow a broad range of applications: With regard to information literacy interventions that primarily focus information-seeking, it can be used to determine baseline levels of information literacy and evaluate training effects through re-administration of the scale. Additionally, it can be used for scientific purposes. For example, it may be used to investigate the role of individual differences (e.g., intelligence or academic self-concept) in the acquisition of information-seeking skills. Furthermore, the scale enables researchers to empirically explore relationships between information-seeking skills and academic achievement, and to compare information-seeking skills across different groups. To sum up, findings from the test can be used both for research as well as to prepare students “for the challenges of living and working in the Information Age.” (Ivanitskaya et al., 2006; Conclusions section, para. 2). Because of its flexible item structure, the test may easily be adapted to other scientific fields by changing a part of the scenario or some approaches: On that account, an adaptation of the PIKE-P for the domain of computer science is currently being developed by the authors.

5 Acknowledgments

Research was funded by a grant from the Leibniz Association, Germany. We would like to thank Prof. Dr. Dirk Wentura from Saarland University for supporting our validation study.

6 References

Artelt, C., Beinicke, A., Schlagmüller, M., & Schneider, W. (2009). Diagnose von Strategiewissen beim Textverstehen [Assessing knowledge about reading strategies]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 41(2),* 96-103.

Andretta, S. (2005). *Information literacy: A practitioner's guide.* Oxford, GB: Chandos Publishing.

Association of College and Research Libraries (2000). *Information literacy standards for higher education.* Retrieved from http://www.ala.org/acrl/files/standards/standards.pdf on 20 May 2014.

Association of College and Research Libraries (2010). *Psychology information literacy standards.* Retrieved from http://connect.ala.org/files/29837/info_lit_psych_pdf_4a60828636.pdf on 20 May 2014.

Association of College and Research Libraries (2013). *Information literacy competency standards for nursing.* Retrieved from http://crln.acrl.org/content/75/1/34.full.pdf on 20 May 2014.

Bowles-Terry, M. (2012). Library instruction and academic success: A mixed-methods assessment of a library instruction program. *Evidence Based Library and Information Practice, 7(1),* 82-95.

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: Analysis of a complex cognitive skill. *Computers in Human Behavior, 21(3),* 487-508.

Dunn, K. (2002). Assessing information literacy skills in the California State University: A progress report. *Journal of Academic Librarianship, 28(1),* 26-35.

Goldhammer, F., Kröhne, U., Keßel, Y., Senkbeil, M., & Ihme, J. M. (2014). Diagnostik von ICT-Literacy. Multiple-Choice- vs. simulationsbasierte Aufgaben [Assessment of ICT literacy: Multiple-choice versus simulation-based tasks]. *Diagnostica, 60(1),* 10-21.

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of Complex Problem Solving: Concept, implementation, and application. *Educational Technology Research & Development*, *61(3),* 407-421.

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),* e6.

Kompetenznetzwerk für Bibliotheken des Deutschen Bibliotheksverbands (2013). Schulungsstatistik Deutschland 2013 [Instruction statistics Germany 2013]. Retrieved from http://www.informationskompetenz.de/fileadmin/DAM/documents/Tabelle_Details_IKde_2013.pdf on 20 May 2014.

Leichner, N., Peter, J., Mayer, A.-K., & Krampen, G. (2013). Assessing information literacy among German psychology students. *Reference Services Review, 41(4)*, 660-674.

Leichner, N., Peter, J., Mayer, A.-K., & Krampen, G. (2014). Assessing information literacy using information search tasks. *Journal of Information Literacy, 8(1),* 3-20.

Macedo-Rouet, M., Rouet, J. F., Ros, C., & Vibert, N. (2012). How do scientists select articles in the PubMed database? An empirical study of criteria and strategies. *European Review of Applied Psychology, 62(2),* 63-72.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). A theoretical basis for situational judgment tests. In J. Weekley & R. Ployhart (Eds.), *Situational judgment tests (pp. 57-81).* San Francisco, CA: Jossey-Bass.

Rosman, T., & Birke, P. (2015). Fachspezifische Erfassung von Recherchekompetenz durch prozedurale Wissenstests: Psychologie vs. Informatik [Discipline-specific assessment of information-seeking ability with procedural knowledge tests: Psychology vs. computer science]. In A.-K. Mayer (Ed.), *Informationskompetenz im Hochschulkontext – Interdisziplinäre Forschungsperspektiven* (pp. 103-120). Lengerich, Germany: Pabst Science Publishers.

Rosman, T., Mayer, A.-K., & Krampen, G. (2014). Combining self-assessments and achievement tests in information literacy assessment: Empirical results and recommendations for practice. *Assessment & Evaluation in Higher Education.* Advance online publication. doi: 10.1080/02602938.2014.950554

Rosman, T., Mayer, A.-K., & Krampen, G. (in press). Intelligence, academic self-concept, and information literacy: the role of adequate perceptions of academic ability in the acquisition of knowledge about information searching. *Information Research.*

Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. Weekley & R. Ployhart (Eds.), *Situational judgment tests (pp. 107-131).* San Francisco, CA: Jossey-Bass.

Sutcliffe, A. G., Ennis, M., & Watkinson, S. J. (2000). Empirical studies of end-user information searching. *Journal of the American Society for Information Science, 51(13),* 1211-1231.

Walsh, A. (2009). Information literacy assessment: Where do we start? *Journal of Librarianship and Information Science, 41(1),* 19-28.

Wentura, D., Ziegler, M., Scheuer, A., Bölte, J., Rammsayer, T., & Salewski, C. (2013). Bundesweite Befragung der Absolventinnen und Absolventen des Jahres 2011 im Studiengang BSc Psychologie [German nationwide survey of bachelor's degree graduates in 2011]. *Psychologische Rundschau, 64(2),* 103-112.

Wise, S. L., Cameron, L., Yang, S. T., Davis, S. L., & Russell, J. (2009). *The Information Literacy Test (ILT): Test Manual.* Harrisonburg, VA: Center for Assessment and Research Studies.

7 Tables

Table 1

Two broader categories enclosing ten sub skills identified in our skill decomposition

Sub skill	**Examples of corresponding ACRL-Outcomes**
Broader Category "Development of Search Strategies"	
1 – Planning behavior (e.g., reasonable sequencing of search steps)	"Defines a realistic overall plan and timeline to acquire the needed information." (ACRL, 2010, p. 3)
2 – Pearl Growing (e.g., searching information based on a relevant publication, the "pearl")	"Selects appropriate subject headings from records of relevant articles to refine search statements." (ACRL, 2013, p. 38)
3 – Extraction of search terms	"Identifies key concepts and terms that describe the information need." (ACRL, 2000, p. 8)
4 – Rewording of search terms	"Identifies keywords, synonyms and related terms for the information needed" (ACRL, 2000, p. 9) and "demonstrates familiarity with the relevant concepts, theoretical perspectives, empirical findings, and historic trends in psychology." (ACRL, 2010, p. 7)
5 – Selection of publication types (e.g., meta-analysis)	"Examines and compares information from various sources." (ACRL, 2010, p. 6)
6 – Selection of search tools (e.g., bibliographic databases)	"… selects the most appropriate sources for *accessing* [emphasis added] the needed information." (ACRL, 2010, p. 3)
Broader Category "Application of Search Strategies"	
7 – Use of Boolean Operators (AND, OR, NOT)	"Creates and uses effective search strategies in relevant databases using … Boolean operators." (ACRL, 2010, p. 3)
8 – Use of online thesaurus	"Uses … controlled vocabulary from the database." (ACRL, 2010, p. 4)
9 – Use of limiters (e.g., by year)	"Uses limiters (e.g., year, population, age, …)." (ACRL, 2013, p. 38)
10 – Full text acquisition (e.g., through the library catalog).	"Retrieves scholarly journals, books, and sources appropriate to the inquiry." (ACRL, 2010, p. 4)

Table 2

Sample item of sub skill 8—Use of online thesaurus

You are searching for longitudinal studies on the effectiveness of cognitive behavior therapy in a reference database. How do you proceed in order to miss as few studies as possible?	not useful at all 1	2	3	4	very useful 5
A) I conduct two searches on the subject headings (thesaurus terms) “cognitive behavior therapy” and “longitudinal studies” and combine these with AND.	□	□	□	□	□
B) I enter “cognitive behavior therapy longitudinal” into the search field.	□	□	□	□	□
C) I conduct a search for the subject heading (thesaurus term) “cognitive behavior therapy”. Then, in the field that contains information about the method used (“Methodology”), I enter “Longitudinal Empirical Study“. Finally, I combine these searches with AND.	□	□	□	□	□
D) I conduct a search for the subject heading (thesaurus term) “longitudinal studies”. Then, in the field that contains information about the research field (“Classification Codes”), I enter “Cognitive Therapy“. Finally, I combine these searches with AND.	□	□	□	□	□

Note. Approach C is the best choice. Conducting a thesaurus search on longitudinal studies (A) is an option, but strongly diminishes the number of hits. Approach D is inappropriate because the combination of a thesaurus search on longitudinal studies (fewer hits) and the limitation on “Cognitive Therapy” (way too broad limitation) entails a very bad hit ratio. Approach B is least appropriate because it does not use advanced database features and will not yield satisfactory results.

Table 3

Sample item of sub skill 6—Selection of search tools

During the writing of your Bachelor thesis, you need several empirical articles about learning strategies of school children aged between 6 and 12 years. How suited are the following tools in order to find the articles?	not useful at all 1	2	3	4	very useful 5
A) Online Library Catalog	□	□	□	□	□
B) Reference database PsycINFO™	□	□	□	□	□
C) Google Scholar™	□	□	□	□	□
D) Reference database PSYNDEX™	□	□	□	□	□

Note. With regard to the rather specific sample (school children aged between 6 and 12 years) and the complex nature of the intended search, bibliographic databases (Approaches B and D) are the best choices. Google Scholar (C) is an option, but its hit ratio on complex searches is reduced. Searching the library catalog (A) is useless, as it only indexes books and is not suited for complex information searches.

Table 4

Sample item of sub skill 4—Rewording of search terms

You are preparing a presentation for a seminar on learning psychology with the following working title: ***"Findings on the efficiency of reward and punishment in toddlers"*** **A search containing the search terms "efficiency", "reward", "punishment", and "toddlers" yielded many irrelevant results. How suited are the following modifications of your search phrase? (Note: Search terms are combined with AND).**	not useful at all 1	2	3	4	very useful 5
A) "Operant conditioning" "toddlers"	□	□	□	□	□
B) "Empirical findings" "efficiency" "reward" "punishment" "toddlers"	□	□	□	□	□
C) "Conditioning" "toddlers" "reinforcement"	□	□	□	□	□
D) "Classical conditioning" "toddlers"	□	□	□	□	□

Note. Approach A (correct scientific term "operant conditioning") is the best alternative. Approach C is slightly less suited because it only contains search terms associated with the correct term. Approaches B and D are useless, because they either do not use a scientific term (B) or use the wrong term (D).

Table 5

Descriptive statistics of the final scale (22 items) for all three samples

	N	***M***	***SD***	**Range**	**Skewness**	**Kurtosis**
Expert study	14	71.42	7.27	58-81	-.38	-1.03
Pilot study	78	53.92	10.41	31-75	-.20	-.33
Validation study	81	53.06	9.82	33-74	-.16	-.73

Note. N = number of subjects; *M* = mean; *SD* = standard deviation; maximum test score = 86.

Table 6

Means, standard deviations and intercorrelations of variables in the validation study

Scale	***M***	***SD***	**Range**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1 PIKE-P	53.06	9.82	33.0-74.0	**(.72)**						
2 ILT-P	21.39	2.31	16.0-26.3	$.51^{***}$	**(.64)**					
3 Full text acquisition tasks	5.01	2.41	0.0-9.0	$.42^{***}$	$.41^{***}$	**(.74)**				
4 Information search tasks—Outcome score	6.54	2.15	1.5-10.0	$.62^{***}$	$.50^{***}$	$.41^{***}$	**(.59)**			
5 Information search tasks—Process score	6.34	2.16	2.5-11.0	$.64^{***}$	$.55^{***}$	$.43^{***}$	$.66^{***}$	**(.75)**		
6 Information search experience	3.09	1.06	1.0-5.0	$.20^{*}$	$.28^{*}$	.16	$.23^{*}$	$.32^{**}$	**(.70)**	
7 Study progress (semesters)	4.30	3.08	1.0-11.0	$.49^{***}$	$.45^{***}$	$.46^{***}$	$.42^{***}$	$.59^{***}$	$.60^{***}$	-

Note. N = 81; PIKE-P = Procedural Information-seeking Knowledge Evaluation—Psychology version; ILT-P = Information Literacy test by Leichner et al. (2013); *M* = mean; *SD* = standard deviation; values in bold on the diagonal = Cronbach's Alpha.

* *p* < .05. ** *p* < .01. *** *p* < .001.

Table 7

Hierarchical Regression Analyses predicting information search task scores from PIKE-P and control variables

	Information search tasks: Outcome rubric			**Information search tasks: Process rubric**		
	β	***R²***	***ΔR²***	***β***	***R²***	***ΔR²***
Block 1		**.37*****			**.45*****	
ILT-P	.34**			.36***		
Information search experience	-.13			-.05		
Study progress	.45***			.46***		
Block 2		**.48*****	**.11*****		**.54*****	**.09*****
ILT-P	.19*			.22*		
Information search experience	-.07			.00		
Study progress	.29**			.31**		
PIKE-P	.40***			.38***		

Note. *N* = 81; Method: Enter; Control variables were entered first (Block 1); PIKE-P was entered subsequently (Block 2); ILT-P = Information Literacy test by Leichner et al. (2013); PIKE-P = Procedural Information-seeking Knowledge Evaluation—Psychology version; *β* = standardized regression weight; R^2 = total variance explained; ΔR^2 = change in R^2 from block 1 to block 2.
* *p* < .05. ** *p* < .01. *** *p* < .001.