Journal of Information Literacy

ISSN 1750-5968

Volume 8 Issue 1
June 2014

Article

Leichner, N., Peter, J., Mayer, A-K. and Krampen, G.. Assessing information literacy programmes using information search tasks. *Journal of Information Literacy*, 8(1) pp. 3-20.
http://dx.doi.org/10.11645/8.1.1870

Assessing information literacy programmes using information search tasks

Nikolas Leichner, Leibniz Institute for Psychology Information, Trier, Germany
Email: mailto:leichner@zpid.de

Johannes Peter, Leibniz Institute for Psychology Information, Trier, Germany
Email: peter@zpid.de

Anne-Kathrin Mayer, Leibniz Institute for Psychology Information, Trier, Germany
Email: mayer@zpid.de

Günter Krampen, Leibniz Institute for Psychology Information, Trier, Germany
Email: krampen@zpid.de

Abstract

The aim of this article is to present information search tasks as an alternative to standardized tests for the assessment of scholarly information literacy (IL). The article describes how a task taxonomy and scoring rubrics were developed as a basis for the construction of standardized search tasks. Based on this taxonomy, sample tasks were created and used in an evaluation study in which an IL instruction programme was scrutinised. In this study, the tasks were applied alongside with a standardized IL test to determine their convergent validity. The results show that IL can be assessed using information search tasks in a reliable and conceptually as well as ecologically valid way. To our knowledge, this is the first publication using information search tasks for the assessment of IL with this degree of standardisation. The task taxonomy, sample tasks, and scoring rubrics are included and can be used by practitioners to create information search tasks tailored to their needs.

Keywords

Information literacy assessment, information seeking behaviour, undergraduate students, Germany.

---

1. Introduction

The concept of information literacy (IL) refers to the competencies to understand when there is a need for information, and subsequently to identify, locate, and evaluate additional information which is required to meet this need (National Forum on Information Literacy n.d.). is considered a crucial skill to manage so-called ‘information overload’ (cf. Hemp 2009) in the workplace and in everyday life (Eisenberg 2008). Scholarly IL (the ability to find and deal with scholarly information, e.g. journal articles) can be seen as a facet of IL. Even though IL is considered a crucial skill for success in higher education (National Forum on Information Literacy n. d.; Johnston and Webber 2003), only few instruments are available to assess this set of skills. The most frequently used data collection method seems to be standardized tests (e.g. Noe and Bishop 2005), notwithstanding that tests lack similarity to authentic situations in the real world (Shavelson 2010). The aim of this article is to present information search tasks and accompanying scoring rubrics that can be used to assess IL in undergraduate psychology students in an ecologically valid fashion as an alternative to the use of standardized tests. The expression ecological validity is used in the sense that ‘the findings obtained within a controlled experiment can be generalised to what is seen in the real

world' (Marcotte and Grant 2010, p. 6). Thus, it is assumed that performance in standardized search tasks (which come close to real-life search tasks) may be interpreted as an indicator of individual competencies highly relevant to information searching tasks in the real world.
The research was funded by the German Joint Initiative for Research and Innovation with a grant acquired in the Leibniz Competition 2012.

2. Literature review

Even though several definitions of IL in higher education have been established (SCONUL 1999; Association of College and Research Libraries 2000; Bundy 2004), a review (Boon et al. 2007) shows that these definitions have in common the following elements:

- The ability to identify when information is needed
- The ability to access and evaluate the accessed information
- The knowledge of ethical issues (e.g. the avoidance of plagiarism).

Due to these similarities, the authors decided to exclusively draw upon the definition by the Association of College and Research Libraries (ACRL 2000), because this framework provides precise performance indicators and has recently been adapted to the field of psychology (ACRL 2010). The definition adapted to psychology includes four standards of IL:

1. Determining the nature and amount of information needed (exemplary performance indicator: 'understands basic research methods and scholarly communication patterns in psychology necessary to select relevant resources');

2. Assessing information effectively and efficiently (exemplary performance indicator: 'selects the most appropriate sources for accessing the needed information');

3. Evaluating information and incorporating information into one's knowledge system (exemplary performance indicator: 'compares new information with prior knowledge to determine its value, contradictions, or other unique characteristics');

4. Using the information effectively to accomplish a specific purpose (exemplary performance indicator: 'applies new and prior information to the planning and creation of a particular project, paper, or presentation').

Because IL is considered an important predictor for achievement in higher education (National Forum on Information Literacy n.d.; Johnston and Webber 2003), and in the workplace (Eisenberg 2008), reliable and valid instruments are needed to measure the corresponding skills, e.g., to determine the performance level of incoming students, or to evaluate IL instruction. Due to their ease of use, IL is often assessed using knowledge tests including multiple choice items, for example, when evaluating IL instruction (Noe and Bishop 2005). At least two multiple choice tests are available commercially (Project SAILS 2013; Center for Assessment and Research Studies 2014). These tests have been shown to provide a reliable and valid way of measuring IL (Center for Assessment and Research Studies 2014).

Other approaches include the analysis of portfolios (Scharf et al. 2007), the analysis of bibliographies that were part of a term paper (Knight 2006), or the use of information search tasks (Julien and Barker 2009); for an overview, see Chang et al. (2012). Integrated approaches based on several instruments have also been developed (cf. Mackey and Jacobson 2007). The approach developed by the California State University (Dunn 2002) for example, included a questionnaire and observation while students were searching for information. For this purpose, screen capture software was used to record activity while searching for information in electronic resources.

Because the students were free to use every resource available (including visiting the library and asking a librarian), they were 'shadowed' by observers while completing the information search

tasks. These approaches can be considered performance assessments (Tung 2010), as they require the students to do more than choosing from several given response options. Performance assessment is seen as a way to assess complex competences instead of factual knowledge. To make the evaluation more consistent, scoring rubrics should be used, as they can guide the scoring process. Rubrics are scoring schemes that help with evaluating the outcome of open-ended tasks. They can be distinguished according to their evaluation criteria, and their specificity. Evaluation criteria can refer to the whole product (holistic rubric), or to several dimensions of the product (analytic rubric). General rubrics can be used to score a broad category of tasks, while specific rubrics are tailored to specific tasks (Moskal 2000; Jonsson and Svingby 2007).

Beside the potential to assess complex competences, the use of rubrics has two more important advantages compared to standardized tests: they make requirements clear, so students know what is expected of them; and they provide evaluators with more detailed information about student learning and performance than standardized tests (Oakleaf 2009). For these reasons, performance assessment including rubrics is becoming more popular in education, although open-ended tasks are by their nature more susceptible to low reliability compared to traditional tests. As a review shows, many rubrics show relatively low reliability; however, reliability can be improved by several actions, e.g. providing benchmarks, or using analytic and specific rubrics (Jonsson and Svingby 2007). Additionally, many problems with rubric assessment seem to be caused by poor rubric design, so they do not seem to be an inherent attribute of rubrics (Oakleaf 2009).

Research has been conducted using less complex and laborious information search tasks (Scott and O'Sullivan 2005; Kim 2009); however, most describe information seeking behaviour instead of evaluating it. Task classifications which have been developed for this purpose do not indicate the difficulty of a certain task type. Ordering tasks by their difficulty, however, is essential for assessing IL using search tasks. Additionally, many studies deal with the analysis of information seeking behaviour while completing non-scholarly search tasks, which differ from scholarly tasks in many respects (e.g. the use of bibliographic databases instead of internet search engines).

3. Purpose of study

Based on the literature review, the authors considered the use of open-ended tasks and scoring rubrics to be a viable way to assess a complex competency, like IL in an ecologically valid fashion. To create several information search tasks with equal structure, a task taxonomy was defined first. The newly designed information search tasks were applied alongside with a standardized IL test which was adopted from previous work by our research group (Leichner et al. 2013). While the test is designed to assess declarative knowledge concerning information retrieval, the search tasks are supposed to measure procedural aspects. These two instruments were used to evaluate an instruction programme for scholarly IL tailored to Psychology undergraduates which combined online and classroom teaching (i.e., used a blended learning approach). Before describing the development of the task taxonomy, a short description the instruction programme and the context in which it was embedded will be given.

3.1 IL instruction at Trier University

The city of Trier is located in the south west of Germany. Compared to other German universities, Trier University is of medium size with a current enrolment of around 15,000 students. The psychology department offers undergraduate and graduate courses (Bachelor and Master degrees); with roughly 1,000 students currently enrolled, it is one of the largest psychology departments in Germany. IL instruction is not part of the regular curriculum; instead, students are expected to acquire these competencies independently. The university library offers corresponding introductory courses; these courses, however, consist of only one 90-minute session and are not tailored to domain-specific information searching problems.

These shortcomings were addressed by a new instruction programme developed as part of a research project conducted at the Leibniz-Centre for Psychology Information (ZPID). The

programme dealt with finding psychology-related information using bibliographic databases (i.e. PsycINFO) and other resources. Further course content included scholarly communication patterns and common publication types in psychology, the use of resources offered by neighbouring disciplines (e.g. educational sciences), and criteria for evaluation of publications (e.g. Journal Impact Factor). The online materials consisted of 10 chapters containing texts, videos and figures which provided information about planning and conducting professional information searches. Through quizzes and exercises at the end of each chapter, students were given the opportunity to test their knowledge and practice information searching. Completing the online modules took about 8 to 10 hours. During the two 90-minutes classroom seminars, participants' questions regarding the online modules were answered, additional exercises were completed, and the content of the course was discussed. The complete instruction programme extended over a period of two weeks.

3.2 Development of the task taxonomy

The first step was therefore the development of a taxonomy that includes three types of scholarly search tasks that have to be solved with the help of electronic resources. The three types of tasks differ with regard to the abilities and competencies required to solve them. As an ability, or competence, knowledge is defined about the existence of information resources (e.g. bibliographic databases like PsycINFO, or scholarly search engines like Google Scholar) and skills concerning the utilisation of these resources. The tasks are designed based on the assumption, that an additive set of competencies is required to solve, i.e. abilities required for type 1 tasks ('easiest' tasks) are also required for type 2 tasks, which, however, require additional abilities. To solve Type 3 tasks, even more abilities are needed to provide an adequate solution. Hence, it is possible to predict that task types may be arranged according to their difficulty. When defining the abilities required for each task, the authors relied on the IL standards provided by the ACRL (Association of College and Research Libraries 2010). The taxonomy which can be used to create new tasks following the same construction pattern is presented in Table 1. The example tasks provided in the table have actually been used during the study. The taxonomy can be used by practitioners to create their own tasks. The following illustration shows how the task type 1 example can be seen as a template that can be adapted: Find two scientific publications published after [enter relevant date] dealing with [enter relevant term]. Use this term as search term.

Table 1: Task taxonomy and sample tasks.

Task type & difficulty level	Description	Competencies required	Example
1	Searching for scientific publications which discuss a topic defined by a scientific term and have been published during a certain period of time.	Range of services of the resources is known and can be used Knowledge of the date filter function of the resource	*Find two scientific publications published after 2005 dealing with short term memory. Use the term „Short Term Memory' as a search term.*
2	Searching for scientific publications discussing an issue which is defined using two scientific terms. Publications must meet several requirements (e.g. publication date, type of methodology used	Understanding of the keyword search function in bibliographic databases, or of the phrase search function in internet resources (i.e. Google Scholar). Understanding of Boolean operators. Understanding of	*Are there meta-analyses published after 2005 investigating 'risk factors' for the development of a 'Posttraumatic stress disorder'? If possible, indicate two publications.*

	in the study).	complex filter functions in bibliographic databases.	
3	Searching for scientific publications concerning an issue which is defined using non-scientific terms.	Development of a complex search strategy (developing search terms, then conducting the actual search). Knowledge of the thesaurus search. Eventually, the use of online dictionaries.	*Find two scientific publications dealing with the question whether differences in the prevalence of mental disorders exist between ethnic groups.*

As IL implies the use of adequate search strategies and resources, as well as finding adequate information, the way the participants approached the tasks (procedure), as well as the outcome (i.e. the references provided) are evaluated based on standardized scoring rubrics (see Appendix B). To conquer the problem of low consistency as far as possible, analytic and task-specific rubrics were used. Criteria were described in great detail to reduce uncertainty in the evaluators. The decision was made to additionally evaluate the procedure because it is possible to find adequate publications by chance without knowledge of adequate resources or precise and systematic search strategies. For example, when working on a type 2 task, it is possible that a participant enters the relevant search term plus 'meta-analysis' in Google Scholar and is able to draw from an extensive list of 'results' two publications meeting all requirements stated in the task description. Using a bibliographic database, which provides a function to filter for certain methodologies, and thus renders more precise results, would indicate a higher level of IL when working on type 2 tasks. When only task outcomes are evaluated, participants might be falsely ascribed a higher level of IL. Additionally, recording the search procedure facilitates comparing results between search tasks differing in the amount of relevant publications ('hits'). If only the outcome was scored, students working on a task with a smaller number of possible 'hits' would be disadvantaged (Kavanagh 2011).

According to the outcome scoring rubric, scores were awarded for every criterion mentioned in the task description (thematic focus of the study, publication date, methodology used) that is met by the publications found. According to the procedure scoring rubric, scores were awarded for choosing an adequate resource (e.g. a bibliographic database), using adequate functions of the resources (filter functions, Boolean operators), and using the thesaurus search (only type 3 tasks). Each facet of this behaviour (e.g. use of Boolean operators; use of filter functions) was scored separately. For example, for a type 2 task, the maximum number of procedure scores was given if the participant worked on the task using bibliographic databases, employing Boolean operators to combine two search terms and limiting the results with the help of the corresponding functions of the database. When developing the procedure rubric, the authors referred to the psychology-specific IL standards (Association of College and Research Libraries 2010) which provide performance indicators for each standard. In that case, reference was made to performance indicators for standard 2; mainly performance indicator 1 (selecting the most appropriate sources) and indicator 2 (constructing an effectively-designed search strategy).

3.3 Hypotheses

The first hypothesis referred to the validity of the taxonomy, particularly the rank order of task difficulties:

1. Those tasks requiring more competencies were expected to be more difficult as revealed by the outcome and procedure scores. Thus, it was expected that type 3 tasks would be more difficult than type 2 tasks, which, in turn, should be more difficult than type 1 tasks.

2. Concerning the relationships among the scores, it was expected that outcome and procedure scores would be moderately correlated, because a sophisticated approach to information searching does not necessarily go hand in hand with good results, and vice versa.
3. Regarding convergent validity, significant correlations between performance on the information search tasks and IL test scores at the baseline time of measurement were expected.
4. Because the training should improve IL, it was expected that the scores on these instruments should be increased by participating in the instruction programme;
5. However, the performance should remain stable when there is no intervention.

4. Method

4.1 Sample

The sample consisted of *N* = 67 undergraduate psychology students who participated in the instruction programme for scholarly IL mentioned above. Of these students, *n* = 34 were in their first year, and *n* = 33 in their second year. On average, the age was 21.67 (*SD* = 2.38); the age range was 18-31. All participants were paid for their participation in the testing sessions. Participants were divided into two groups. The first group consisted of *n* = 37 participants, the second group of *n* = 30. The two groups did not differ in their baseline IL levels at the first time of measurement.

4.2 Measures and Procedure

The duration of the study was four weeks, while the actual training took only two weeks. Performance levels were assessed three times during this period in the computer lab of Trier University. Measurement 1 (t1) took place right at the start of the study. Afterwards, the first group completed the instruction programme while the second group served as a waiting control group. After two weeks performance was assessed for the second time (t2). After that, the second group completed the programme. The final measurement (t3) took place after all participants had completed the instruction programme. Assessments were conducted using survey software which presented the tasks and questionnaires and collected the answers. Participants took part in the assessment sessions in groups of 15 to 20 participants. To minimise distortion of the results, it was made sure that the participants would not interact with each other, as in an exam situation. The participants did not receive any help from the instructor during the assessment (except to solve technical problems with the survey software).

Based on the task taxonomy, three information search tasks of every type were created. The complete list of tasks can be found in Appendix A. At each time of measurement, participants completed, among other measures, three information search tasks (one task of each type) followed by a standardized IL test. The test was adapted from prior work of our research group (Leichner et al., 2013) and contained 35 multiple choice items. When completing the information search tasks, the participants were presented with one task of each type; time allowed for the completion of each task was limited. The tasks were presented to the participants ordered by difficulty. Participants were not allowed to skip one task and return to it later. To complete the information search tasks, participants could use all resources available on the computers in the lab, which are access to the internet, to bibliographic databases, and to online library catalogues. Participants had to record the publications they had found by entering the bibliographic information into input boxes that were provided by the survey software. Upon completion of each task, the participants were asked several questions concerning their approach to the search procedure (e.g. the search engine/bibliographic database used, search terms used, filter functions used). This data was used to score the procedure.

5. Results

All answers were scored independently by two raters and each task was evaluated separately. To examine whether the answers can be rated reliably on the basis of the scoring rubric (i.e. different raters award equal scores to the same answer), interrater-reliability coefficients (correlation between the scores awarded by the two raters) were calculated. According to the literature, the interrater-reliability coefficients should be at least $r = .60$ (Hartmann 1977). In our study, the coefficients for the outcome scores ranged from $r = .62$ to $r = .87$; most correlation coefficients were above .70. For the procedure scores, the inter-rater correlation for each facet (e.g. search engine used) ranged from $r = .70$ to $r = .92$. When the scores diverged, both raters agreed on one judgment which was used for analysis.

First, it was determined whether the expected order of task difficulties corresponded to the empirical data (hypothesis 1). The scores awarded for the information search tasks were scaled to restrict their range from 0 to 1 (ratio of observed score and maximum score per task). Thus, the values in table 2 can be multiplied with 100 to obtain the percentage of the maximum scores that was actually achieved on average. As expected, type 1 tasks were easier than type 2 tasks, which, in turn, were easier than type 3 tasks. The scores are displayed in table 2; t2 is left out to provide a better overview of the results.

Table 2: Mean scores (standard deviations in brackets) for search task outcome and procedure ordered by task type before and after participation in the instruction programme

	Task type 1	Task type 2	Task type 3
Outcome t1	0.77 (0.28)	0.50 (0.30)	0.32 (0.26)
Outcome t3	0.91 (0.20)	0.87 (0.18)	0.60 (0.36)
Procedure t1	0.55 (0.19)	0.46 (0.17)	0.36 (0.16)
Procedure t3	0.79 (0.11)	0.72 (0.15)	0.74 (0.18)

Remarks: outcome = scores awarded for the adequacy of publications found; procedure = scores awarded for the search behaviour; t1/t3 = time of measurement 1, respectively 3.

Repeated measures analyses of variance were computed to examine whether differences in achieved scores between tasks at one time of measurement were of statistical significance. These analyses revealed significant differences between the tasks concerning the outcome scores ($F[2,132] = 45.66$, $p < .01$) and procedure scores ($F[2,132] = 29.65$, $p < .01$) at t1. Linear contrasts showed that all differences (i.e. the difference between task type 1 and 2, and task type 2 and 3, respectively) reached significance level. At t3, analyses of variance again showed significant differences between the tasks on the outcome scores ($F[2,132] = 29.26$, $p < .01$), and the procedure scores ($F[2,132] = 5.01$, $p < .01$). Linear contrasts revealed that the difference on the outcome scores between tasks 1 and 2 and the difference on the procedure scores between tasks 2 and 3 did not reach significance, even though tendencies remain visible.

For the following analyses, scores of all three information search tasks presented at one time of measurement were averaged separately regarding the outcome and procedure scores, so that two scores for each time of measurement (outcome score and procedure score) emerged. The means of these scores are displayed in figures 1 and 2. Stability coefficients for the scores could be computed as participants from group 2 participated in the assessment sessions at t1 and t2 without intervention in the meantime. The stability coefficients were computed by correlating the scores obtained at t1 and t2. They indicate whether the scores are stable over time what is considered a quality characteristic of an instrument. The correlation coefficient was $r = .53$ for the outcome scores, and $r = .63$ for the procedure scores. For the standardized IL test, the correlation coefficient was $r = .71$.

Figure 1: Scores for the search task outcome with standard errors

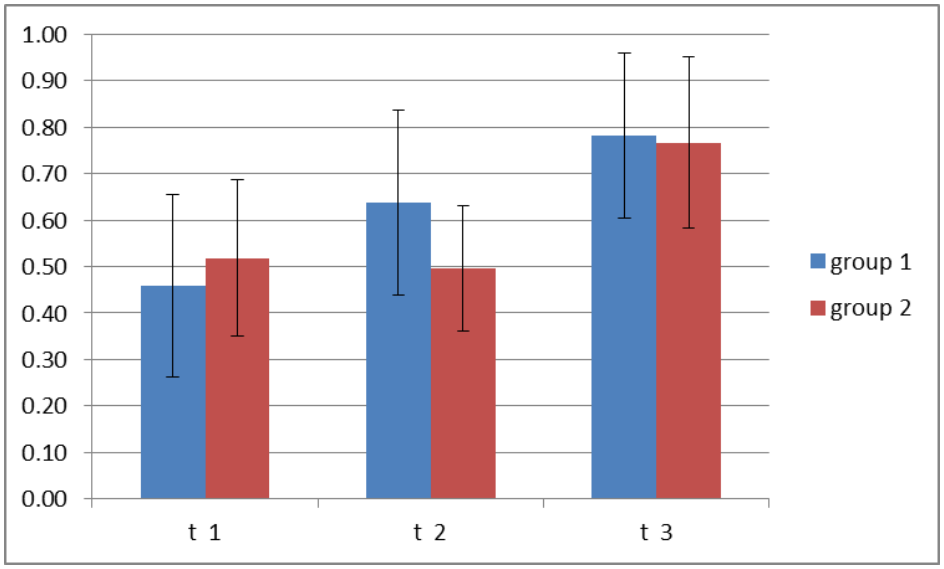
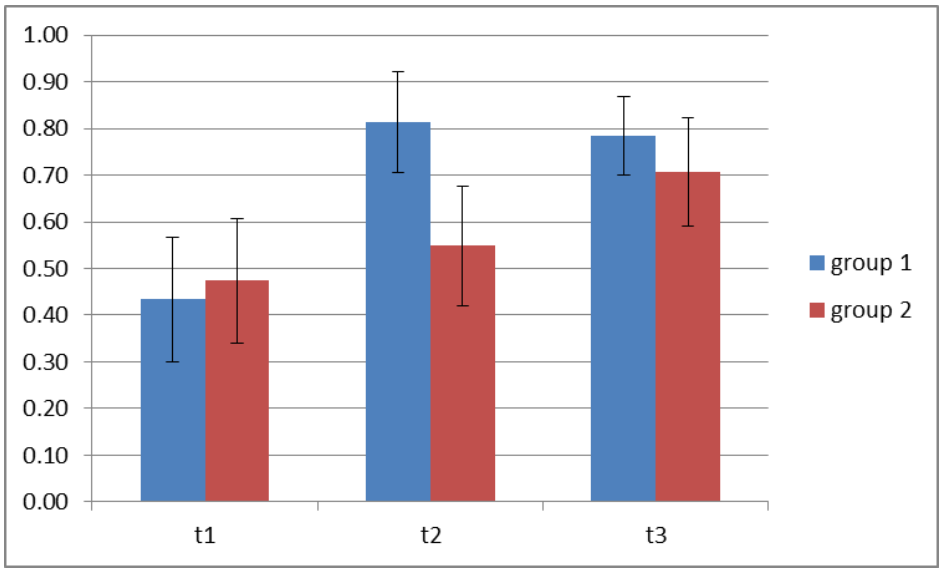


Figure 2: Scores for the search task procedure with standard errors



Correlation analyses with regard to hypothesis 2 revealed that there was a significant relationship between outcome and procedure scores at the first time of measurement ($r = 0.22$, $p < 0.05$). Significant correlations between the two measures were also found at t2 ($r = .41$, $p < .01$) and t3 ($r = .32$, $p < .01$).To validate the information search tasks, correlations between the IL test and outcome and procedure scores at t1 (see hypothesis 3) were computed. It was found that there was a significant correlation between the IL test and the outcome score ($r = 0.29$, $p < 0.01$), and the procedure score ($r = 0.48$, $p < 0.01$).

5.1 Evaluation results

As mentioned in the sample section, the two groups did not differ in their baseline performance at t1 ($t[65] = 1.34$, n.s. for the outcome scores, $t[65] = 1.23$, n.s. for the procedure scores, and $t[65] = 0.78$, n.s. for the IL test). As expected, the two groups differed at t2 ($t[65] = 3.32$, $p < .01$, and $t[65] = 9.21$, $p < .01$, respectively) with group 1 outperforming group 2. At t3, the groups did not differ on their outcome scores ($t[65] = 0.32$, n.s.), but on their procedure scores ($t[65] = 3.21$, $p < .01$). On the procedure scores, group 1 performed better than group 2 at t3.

The next step in the evaluation of the IL instruction programme was to compute two separate repeated measures analyses of variance (ANOVA) with time of measurement as within subjects and group membership as between subjects independent variables. Search outcome and procedure respectively were used as dependent variables. These analyses examined the patterns of change of the two groups. Both analyses revealed a significant interaction of the two independent variables (for the outcome scores $F[2,130] = 5.45$, $p < .01$, and for the procedure scores $F[2,130] = 37.38$, $p < .01$). This indicates that mean scores of the two groups changed differently.

To analyse these differential changes in more detail, separate analyses for the two groups were computed. When analysing the outcome scores, a significant main effect for time of measurement (within subjects) for group 1 ($F[2,72] = 27.17$, $p < 0.01$) was found, as well as for group 2 ($F[2,58]= 25.63$, $p < 0.01$). For group 1, linear contrasts revealed that there were significant differences between all times of measurement. That is, group 1 participants (which had participated in the training) improved from t1 to t2, and from t2 to t3. For group 2, there was no significant difference between t1 and t2, but between t2 and t3, so these participants improved only from t2 to t3 after having completed the instruction programme.

Analysing the procedure scores, the main effect for time of measurement also reached significance for both groups ($F[2,72] = 130.55$, $p < 0.01$ and $F[2,58] = 54.31$, $p < 0.01$, respectively). For group 1, linear contrasts revealed a significant difference between t1 and t2, while the difference between t2 and t3 was not significant. For group 2, linear contrasts revealed significant differences between all three times of measurement. That is, group 1 participants improved from t1 to t2; while group 2 participants improved between all three time of measurement.

6. Discussion

In the present study, a taxonomy of scholarly information search tasks in the domain of psychology (including scoring rubrics for search procedures and search outcomes) was developed and tested empirically. With regard to the reliability of scoring, inter-rater reliability coefficients were computed. The correlations between the two raters can be considered appropriate as they all exceeded the minimum value of $r = .60$ for interrater-reliability (Hartmann 1977; Gelfand and Hartmann 1984). It is obvious that interrater-reliabilities are higher for the procedure scores than for the outcome scores, as the judgment whether an article covers a certain topic is more ambiguous than rating whether a certain resource has been used or not. The stability coefficients for both scores were in an acceptable range (cf. March et al. 1999). Taken together, these findings indicate that it is possible to create rubrics for scoring scholarly information search tasks meeting the requirements of reliable testing. Of course, the standardized test showed a higher stability coefficient what is in accordance with the existing literature.

The hypotheses regarding the validity of the measures could predominantly be verified. The expected order of task difficulties was supported empirically. This indicates that tasks requiring more competencies are more difficult for participants confirming the assumptions made when designing the task taxonomy. Differences in task difficulty are less marked at t3 due to increased overall performance levels and to ceiling effects in some cases. A common criterion for the detection of ceiling effects is that several participants reach the maximum score, while the overall performance level is very high. In line with our expectations, information search tasks outcome and procedure correlated significantly, even though correlations are only low to moderate. This may be due to the fact that using the ‘most adequate’ procedure does not necessarily lead to good results.

A moderate but significant correlation between performance on the IL test and the search tasks could be found, which may be interpreted as a hint to the convergent validity of the search tasks. As the test is supposed to measure declarative knowledge related to information retrieval, while the search tasks are intended to measure procedural aspects, the moderate correlations are conforming to our expectations. Obviously, the two instruments capture different facets of information retrieval skills. As the information search tasks have more similarity with real-world

tasks than the standardized test, it is reasonable to assume that they capture the ability to solve scholarly information problems better than a standardized test. Proponents of performance assessment (Oakleaf 2009; Tung 2010) would certainly endorse this assumption.

When comparing search task performance before and after the instruction programme, the authors found that subjects' scores were higher on both search task evaluation variables after participation, a finding that is in line with expectations. Firstly, it demonstrates that the instruction programme was effective. Secondly, it can also be interpreted as a piece of evidence for the validity of the information search tasks. It was also assumed that performance would not change when there is no intervention; however, some of our empirical findings (e.g. the significant difference on the procedure scores of group 2 between t1 and t2) are not in line with this hypothesis. To explain these changes in performance without intervention, it seems plausible to refer to testing effects (an increase of performance due to experience with the type of tasks used in our study, see Kulik et al. 1984; Hausknecht et al. 2007). It is also possible, however, that the participant's interest in the topic was stimulated by taking the test, so the participants plunged into the domain of information search and improved their skills on their own.

Another noticeable finding is that outcome and procedure scores developed differently over the course of the study. Apart from the most obvious explanation after which participants with poor information seeking behaviour found publications by chance, another possibility should be considered. It may be questioned whether participants indicating the use of sophisticated information resources (e.g. the use of bibliographic databases and the use of complex filter functions provided by the databases) were able to use these resources in an efficient way. It seems plausible that many participants knew of these information resources and used these resources without knowing how to utilise them effectively. This assertion shows a parallel to Miller's pyramid of competence (Miller 1990) which has been used to determine the nature of tests e.g. (Wass et al. 2001). Miller claims that a person with (declarative) knowledge does not necessarily know how to apply this knowledge.

Several limitations of our efforts should not be withheld. The first one is the fact that the two groups participated in the instruction programme during different periods of time. However, as linear contrasts revealed that both groups improved significantly through participation, this limitation can impair the weight of the findings only minimally. Second, regarding the ecological validity of our study, it can be called into question whether the exam situation which was staged at each time of measurement permits ecologically valid assessment at all. The authors do not think that this is a problem as no grades were assigned. The participants were explicitly told that the assessment was part of a scientific study and that their scores would not be used for any other purpose. The authors cannot, however, rule out completely that student performance was influenced by these circumstances. Though, any attempt to make the assessment more similar to the real world would have undermined standardisation. Third, because the information search tasks are designed as an ecologically valid indicator of IL, they are more resource-intensive than other standardized instruments. Because scoring of the tasks has to be conducted manually, it requires more time than the automatized calculation of test scores. So, the approach suggested here might be of limited use in practice.

Future research might aim at developing less laborious alternatives that can be used to replace the information search tasks. A fruitful approach might be to ask the participant to put himself into the position of a student who has to find literature. Then, the participant is given several options (e.g. use of Google Scholar, bibliographic database, library catalogue) and has to indicate which seems to be most useful to him being in that situation. Further research might also deal with the different developments of search task outcome and procedure scores over the course of the study.

7. Conclusion

To sum up, the taxonomy of search tasks presented in this paper seems to be a good basis for the creation of scholarly information search tasks. The taxonomy and the scoring rubrics can be used

by practitioners to create their own information search tasks. The authors were also able to show that information search tasks based on this taxonomy are a reliable and valid instrument to assess IL in college students. Hence, the information search tasks will be used in further scientific studies at our research institute. For example, this instrument has already been used to estimate the convergent validity of a newly-developed IL test.
This is in line with authors arguing that rubrics are suitable for reliable assessment (Oakleaf 2009). The information search tasks also come much closer to real-life tasks than multiple choice instruments and follow a call for hands-on assessment tasks made by several authors (Shavelson 2010). For example, the authors believe their information search tasks come close to the information searches performed by psychology students when preparing a term paper.

Another point worth mentioning is that this design gives interesting insights into how students use electronic resources. For example, the authors were quite surprised to find out that a significant number of students already used bibliographic databases before participating in the instruction programme. However, because they did not use the complex functions offered by bibliographic databases (e.g., the thesaurus or the option to search for publications applying a specific methodology), these students failed to reach high scores on the search tasks. These findings suggest that the skills necessary to perform advanced information searches are not acquired independently by psychology students but they can be developed during information literacy courses.

References

Association of College and Research Libraries. 2000. *Information literacy competency standards for higher education* [Online]. Chicago, IL: American Library Association. Available at: http://www.ala.org/acrl/sites/ala.org.acrl/files/content/standards/standards.pdf [Accessed 6 June 2012].

Association of College and Research Libraries. 2010. *Psychology information literacy standards* [Online]. Chicago, IL: American Library Association. Available at: http://www.ala.org/acrl/standards/psych_info_lit [Accessed 19 March 2013].

Boon, S. et al, 2007. A phenomenographic study of English faculty's conceptions of information literacy. *Journal of Documentation* 63(2), pp. 204–228. http://dx.doi.org/10.1108/00220410710737187

Bundy, A., ed. 2004. *Australian and New Zealand information literacy framework: principles, standards, and practice.* 2nd ed. Adelaide: Australian and New Zealand Institute for Information Literacy .

Center for Assessment and Research Studies. 2014. *Information Literacy Test (ILT)* [Online]. Boulder, CO: Madison Assessment. Available at: http://www.madisonassessment.com/assessment-testing/information-literacy-test [Accessed 20 February 2014].

Chang, Y.-K et al, 2012. Assessing students' information literacy skills in two secondary schools in Singapore. *Journal of Information Literacy* [Online] 6(2), pp. 19–34. http://dx.doi.org/10.11645/6.2.1694.

Dunn, K. 2002. Assessing information literacy skills in the California State University: a progress report. *Journal of Academic Librarianship* 28(1), pp. 26–35.
http://dx.doi.org/10.11645/6.2.1694
Eisenberg, M.B. 2008. Information literacy: essential skills for the information age. *DESIDOC Journal Of Library & Information Technology* [Online] 28(2), pp. 39–47. Available at: http://publications.drdo.gov.in/ojs/index.php/djlit/article/view/166/77 [Accessed 29 January 2014].

Gelfand, D.M. and Hartmann, D.P. 1984. *Child behavior analysis and therapy.* 2nd ed. New York: Pergamon Press.

Hartmann, D.P. 1977. Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis* 10, pp. 103–116. http://dx.doi.org/10.1901/jaba.1977.10-103

Hausknecht, J.P. et al, 2007. Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology* 92(2), pp. 373–385. http://dx.doi.org/10.1037/0021-9010.92.2.373

Hemp, P. 2009. Death by information overload. *Harvard Business Review* [Online] September 2009, pp. 83–89. Available at: http://hbr.org/2009/09/death-by-information-overload/ar/1b [Accessed 30 January 2014].

Johnston, B. and Webber, S. 2003. Information literacy in higher education: a review and case study. *Studies in Higher Education* 28(3), pp. 335–352. http://dx.doi.org/10.1080/03075070309295

Jonsson, A. and Svingby, G. 2007. The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review* [Online] 2(2), pp. 130–144. http://dx.doi.org/10.1016/j.edurev.2007.05.002

Julien, H. and Barker, S. 2009. How high-school students find and evaluate scientific information: a basis for information literacy skills development. *Library & Information Science Research* 31(1), pp. 12–17. http://dx.doi.org/10.1016/j.lisr.2008.10.008

Kavanagh, A. 2011. The evolution of an embedded information literacy module: using student feedback and the research literature to improve student performance. *Journal of Information Literacy* [Online] 5(1), pp. 5–22. Available at: http://ojs.lboro.ac.uk/ojs/index.php/JIL/article/view/LLC-V5-I1-2011-1 [Accessed 24 February 2014].

Kim, J. 2009. Describing and predicting information-seeking behavior on the Web. *Journal of the American Society for Information Science and Technology* 60(4), pp. 679–693. http://dx.doi.org/10.1002/asi.21035

Knight, L.A. 2006. Using rubrics to assess information literacy. *Reference Services Review* 34(1), pp. 43–55. http://dx.doi.org/10.1108/00907320610640752

Kulik, J.A et al, 1984. Effects of practice on aptitude and achievement test scores. *American Educational Research Journal* 21(2), pp. 435–447. http://dx.doi.org/10.3102/00028312021002435

Leichner, N. et al, 2013. Assessing information literacy among German psychology students. *Reference Services Review* 41(4), pp. 660–674. http://dx.doi.org/10.1108/RSR-11-2012-0076

Mackey, T.P. and Jacobson, T. 2007. Developing an integrated strategy for information literacy assessment in general education. *The Journal of General Education* 56(2), pp. 93–104. http://dx.doi.org/10.1353/jge.2007.0021

March, J.S. et al, 1999. Test-retest reliability of the multidimensional anxiety scale for children. *Journal of Anxiety Disorders* [Online] 13(4), pp. 349–358. http://dx.doi.org/10.1016/S0887-6185(99)00009-2

Marcotte, T.D. and Grant, I. 2010. *Neuropsychology of everyday functioning*. New York: Guilford Press.

Miller, G.E. 1990. The assessment of clinical skills/competence/performance. *Academic medicine* 65(9), pp. S63-67. http://dx.doi.org/10.1097/00001888-199009000-00045

Moskal, B.M. 2000. Scoring rubrics: what, when and how? *Practical Assessment, Research & Evaluation* [Online] 7(3). Available at: http://PAREonline.net/getvn.asp?v=7&n=3 [Accessed 30 January 2014].

National Forum on Information Literacy. n. d. *What is information literacy?* [Online]. Available at: http://infolit.org/?page_id=3172 [Accessed 6 June 2012].

Noe, N.W. and Bishop, B.A. 2005. Assessing Auburn University Library's tiger information literacy tutorial (TILT). *Reference Services Review* 33(2), pp. 173–187. http://dx.doi.org/10.1108/00907320510597372

Oakleaf, M. 2009. Using rubrics to assess information literacy: an examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology* 60(5), pp. 969–983. http://dx.doi.org/10.1002/asi.21030

Project SAILS. 2013. *Project SAILS information literacy assessment* [Online]. Kent, OH: Kent State University. Available at: https://www.projectsails.org/ [Accessed 16 August 2013].

Scharf, D. et al, 2007. Direct assessment of information literacy using writing portfolios. *Journal of Academic Librarianship* 33(4), pp. 462–477. http://dx.doi.org/10.1016/j.acalib.2007.03.005

SCONUL 1999. *Information skills in higher education*. London, UK: SCONUL

Scott, T.J. and O'Sullivan, M.K. 2005. Analyzing student search strategies: making a case for integrating information literacy skills into the curriculum. *Teacher Librarian* 33(1), pp. 21–25.

Shavelson, R.J. 2010. On the measurement of competency. *Empirical Research in Vocational Education and Training* 2(1), pp. 41–63.

Tung, R. 2010. *Including performance assessments in accountability systems: a review of scale-up efforts*. Boston, MA: Center for Collaborative Education.

Wass, V., Van der Vleuten, C., Shatzer, J. and Jones, R. 2001. Assessment of clinical competence. *Lancet* 357(9260), pp. 945–949. http://dx.doi.org/10.1016/S0140-6736(00)04221-5

Appendix A: Information search tasks

The instruction was presented at every instance of measurement. After that, three search tasks were given in the order shown here. A type 1 task was presented first, followed by a type 2 then a type 3 task.

Instruction

You are working as a graduate assistant for the psychology department of your university. A part of your responsibilities is to find literature on certain topics. In the following, one of the professors will assign you several tasks. The tasks have to be completed within a time limit. Please write down:

1. The author of the publication (the first author only)
2. Year of publication
3. Title of the publication.

After the completion of each task, you will be asked several questions concerning your approach to the task.

Time of measurement 1

1. The professor wants to find out whether a specific issue has been the topic of recent publications. Your assignment is: find two scientific publications published after 2005 dealing with false memories. Use the term ‘false memory’ as a search term.
2. The professor wants to gain an overview about a specific issue. Your assignment is: are there longitudinal studies published after 2005 investigating ‘risk factors for generalized anxiety disorder’? If possible, indicate two publications.
3. The professor wants to find out whether the treatment of depressive disorders is more difficult in older persons. Your assignment is: find two scientific publications dealing with the question whether age differences have an impact on the treatment outcome of a depression treatment.

Time of measurement 2

1. The professor wants to find out whether a specific issue has been the topic of recent publications. Your assignment is: find two scientific publications published after 2005 dealing with short term memory. Use ‘Short Term Memory’ as a search term.
2. The professor wants to gain an overview about a specific issue. Your assignment is: are there meta-analyses published after 2005 investigating ‘risk factors’ for the development of a ‘Posttraumatic stress disorder’? If possible, indicate two publications.
3. The professor wants to find out whether mental disorders tend to occur more often among members of certain ethnic groups. Your assignment is: find two scientific publications dealing with the question whether differences in the prevalence of mental disorders exist between ethnic groups.

Time of measurement 3

1. The professor wants to find out whether a specific issue has been the topic of recent publications. Your assignment is: find two scientific publications published after 2003 dealing with panic attacks. Use the term ‘panic attack’ as a search term.
2. The professor wants to gain an overview about a specific issue. Your assignment is: are there meta-analyses published after 2003 investigating the effectiveness of ‘cognitive behavior therapy’ for the treatment of depression (‘Major depression’)? If possible, indicate two publications.

3. The professor tries to find out which pain treatment options are available for patients with personality disorders. Your assignment is: find two scientific publications dealing with the questions which kinds of pain treatment can be used with patients with personality disorders.

List of relevant search terms for the type 3 tasks

This list was not presented to participants, but was used to score the tasks, as described below:

1. Time of measurement 1: age differences; major depression; treatment outcomes, or sychotherapeutic outcomes, or treatment effectiveness evaluation.
2. Time of measurement 2: racial and ethnic differences; epidemiology; mental disorders
3. Time of measurement 3: personality disorders; pain management

Questions concerning the approach taken

The following questions were asked after the completion of every task:

- Which search engine/ bibliographic databases did you use to solve the task? (Please indicate the search engine used last only)
- Which search terms did you use?
- How did you determine the search which you used? (asked after type 3 tasks only)
- Did you use any specific functions of the resource (e.g. filters)? If yes, please indicate which functions have been used.
- Did you use any additional search engines/bibliographic databases? If yes, please indicate which additional resources have been used.
- Did these resources contribute to the outcome?

Appendix B: Rubrics

Outcome rubric

Type 1 tasks

Every publication indicated was evaluated separately:

- Thematic relevance: the search term given in the task description (or a synonym) is included in the title of the publication or the abstract: 0.5 points, else 0 points.
- Publication date: the publication had been published during the timeframe defined in the task description: 0.5 points, else 0 points.
-

Per publication, a maximum score of one is possible. For the whole task, a maximum of two points can be awarded.

Type 2 tasks

Every publication indicated was evaluated separately:

- Thematic relevance: the search term given in the task description (or a synonym) is included in the title of the publication, the abstract, or the subject headings: 0.5 points per term, else 0 points. As the tasks require the use of two terms, a maximum of 1 point can be awarded.
- In case, the title reveals that the publication does not deal with the question, although the publication is ascribed both subject headings in a bibliographic database, 0.5 points were awarded.
- Publication date: the publication had been published during the timeframe defined in the task description: 0.5 points, else 0 points.
- Methodology used: the publication uses the methodology indicated in the task description (e.g. meta-analysis): 0.5 points, else 0 points.

Per publication, a maximum score of two is possible. For the whole task, a maximum of four points can be awarded.

Type 3 tasks

Every publication indicated was evaluated separately.

- Thematic relevance: The relevant search terms (see list), or synonyms of these terms, are contained in the title, abstract, or subject headings: 2 points.
- Only one of the relevant terms is contained in the title, abstract, or subject headings: 0.5 points

Per publication, a maximum score of two is possible. For the whole task, a maximum of four points can be awarded.

Procedure rubric

This rubric can be used to score the answers given by the participants to the questions defined in section 1.6.

Type 1 tasks

- Search engine: use of Google Scholar, MS Academic Search, or bibliographic databases: 1 point, else 0 points.

- Search terms: search term given in the task description has been used: 1 point, else 0 points.
- Specific functions: filter for the publication year has been used: 1 point, else 0 points.
- Use of additional resources: use of Google Scholar, MS Academic Search, or bibliographic databases: 1 point each, else 0 points.
- Contribution of the additional resources: not evaluated; data has been collected for descriptive purposes only.

No more than four points were awarded per task.

Type 2 tasks

- Search engine: bibliographic databases: 1 point; Google Scholar, or MS Academic Search: 0.5 points, else 0 points.
- Search terms: search terms given in the task description have been used: 1 point, else 0 points.
- Specific functions: use of Boolean operators, filter for publication year, or methodology used, and every additional function which was required to solve the tasks efficiently: 1 point for each function.
- Use of additional resources: use of Google Scholar, MS Academic Search, or bibliographic databases: 0.5 points each, else 0 points.
- Contribution of the additional resources: not evaluated; data has been collected for descriptive purposes only.

No more than five points were awarded per task.

Type 3 tasks

- Search engine: bibliographic databases: 1 point; Google Scholar, or MS Academic Search: 0.5 points, else 0 points.
- Search terms: one point for every relevant term (see list above); 0.5 points for each similar term; else 0 points.
- How did the participant generate the search terms? Extracting from existing literature[1] or thesaurus search provided by a bibliographic database: 1 point, else 0 points.
- Specific functions: not evaluated; data has been collected for descriptive purposes only.
- Use of additional resources: Google Scholar, MS Academic search, or bibliographic databases: 0.5 points each, else 0 points.
- Contribution of the additional resources: not evaluated; data has been collected for descriptive purposes only.
-

No more than five points were awarded per task.

Information about how to use the tasks and rubrics:

Time for the completion of the search tasks was limited. Five minutes were given for tasks of type one; ten minutes for type two tasks and 15 minutes for type three tasks.

When evaluating the appropriateness of the publications found, we did not refer to the full text of the publications indicated by the participants. Due to time constraints, the evaluation is solely based on their titles of the publications and the abstracts. If a publication was definitely not scientific (e.g. fiction), 0 points were awarded across the board.

Scoring the tasks as described above will take about 15 minutes per participant for each time of measurement.

Notes

The participant might find an article using Google, Google Scholar, or any other internet resource. The search terms can be extracted from this article, or the participant might look up the terms ascribed to this publication in a bibliographic database.