Journal of Information Literacy

ISSN 1750-5968

Volume 4 Issue 1

June 2010

Article

Helvoort J. van 2010. A scoring rubric for performance assessment of information literacy in Dutch Higher Education. *Journal of Information Literacy*, 4(1), pp. 22-39 *http://ojs.lboro.ac.uk/ojs/index.php/JIL/article/view/PRA-V4-I1-2010-2*

A scoring rubric for performance assessment of information literacy in Dutch Higher Education

Jos van Helvoort, Senior Lecturer, The Hague University of Applied Sciences
Email: a.a.j.vanhelvoort@hhs.nl

Abstract

The main purpose of the research was the development and testing of an assessment tool for the grading of Dutch students' performance in information problem solving during their study tasks. Scholarly literature suggests that an analytical scoring rubric would be a good tool for this. Described in this article are the construction process of such a scoring rubric and the evaluation of the prototype based on the assessment of its usefulness in educational practice, the efficiency in use and the reliability of the rubric. To test this last point, the rubric was used by two professors when they graded the same set of student products. 'Interrater reliability' for the professors' gradings was estimated by calculating absolute agreement of the scores, adjacent agreement and decision consistency. An English version of the scoring rubric has been added to this journal article as an appendix. This rubric can be used in various discipline-based courses in Higher Education in which information problem solving is one of the learning activities. After evaluating the prototype it was concluded that the rubric is particularly useful to graders as it keeps them focussed on relevant aspects during the grading process. If the rubric is used for summative evaluation of credit bearing student work, it is strongly recommended to use the scoring scheme as a whole and to let the grading work be done by at least two different markers.

Keywords
Information Literacy; Performance Assessment; Scoring Rubrics; Higher Education; Netherlands

1. Introduction

In recent decades Dutch Higher Education institutes, particularly in the Applied Sciences disciplines, have shifted their educational strategy from 'teacher-centred' instruction to 'student-centred' learning (Tigelaar et al. 2004, p. 254). The new pedagogical systems are known as 'Problem based learning', 'Project based learning' and 'Competence oriented learning' and they all use the principle that students create their own knowledge during their learning experiences in line with the 'constructivist learning theory' (Brand-Gruwel et al. 2005, p. 488; Dochy et al. 2005, p. 43; Elshout-Mohr et al. 2002, p. 370). A key factor in these learning experiences is the availability of Powerful Learning Environments (PLEs) that are rich in resources and learning materials to enable students' personal exploration of the problems or student tasks (Dochy and McDowell 1997, p. 283). Nowadays it is clear that curricula in the Higher Education system in the Netherlands strongly encourage students to solve their own information problems. However, as Brand-Gruwel et al. (2005, p. 488) remarked, it cannot be assumed that students have acquired these skills by themselves, which is why information specialists and educational reformers often plead for information literacy instruction in the Higher Education curricula (see for instance: Dirkx et al. 2006, p. 19; Walraven et al. 2008, p. 624). To be effective, this instruction should not be limited to information searching or retrieval but should include the recognition of information needs, the evaluation of information and the processes of information use and dissemination (Boekhorst 2003, p. 299), in other words it should cover the complete process of information problem solving (Brand-Gruwel et al. 2005, p. 490).

However, although there is an increasing focus of attention on information literacy in Higher Education institutes, it appears that students very often do not recognise the importance of this competence (Drent and Timmers 2006, p. 1; Gross and Latham 2007, p. 336). The reason for this might be that the students' information seeking processes and the way in which they use the information they have gathered, are hardly ever included in the grading of the assignments in their discipline based courses. It is obvious that students do not pay much attention to the process of information problem solving if it is not graded, because the assessment requirements strongly influence their learning behaviour (Driessen and Van der Vleuten 2000, p. 236; Knight 2006, p. 52; Stubbings and Franklin 2006, p. 7). That is why I decided to start a research project on the construction of an assessment tool for credit bearing performance assessment of information literacy. According to Nitko and Brookhart (2007, p. 244) a performance assessment "(a) presents a task requiring students to do an activity that requires applying their knowledge and skills from several learning targets and (b) uses clearly defined criteria to evaluate how well the student has achieved this application". My assumption was that the use of such a credit bearing performance assessment would increase students' interest in obtaining good information processing skills, because students appear to have the tendency to "be assessment driven and behave in a strategic way" (Cochrane 2006, p. 105).

2. Literature review

The review of the literature focuses on the instruments that can be used in Higher Education for a performance assessment of information literacy skills. When I started my research in November 2006 little had been written on this topic. Gratch-Lindauer (2003) describes several types of assessment tools for measuring information literacy learning outcomes: tests, quizzes, embedded course assignments, direct observation, questionnaires, research diaries and portfolios. In her overview she uses a classification of three learning domains that goes back to Bloom's taxonomy of learning objectives:

- Cognitive (what do students know?).

- Behavioural – performance based (what can students do?). This domain refers to what Bloom mentions as the "psychomotor domain" (Bloom 1956, p. 7).
- Affective (how do students perceive their abilities?).

Gratch-Lindauer notes that most assessment instruments fit well into a specific learning domain (tests and quizzes in the cognitive domain, embedded assignments, direct observation and portfolios in the behavioural domain and questionnaires and research diaries in the affective domain) but she emphasises that most tools can be used for at least two different dimensions of the information literacy learning outcomes. In her view quizzes and tests, for instance, may include items that refer to the behavioural domain (pp. 27-28).

2.1 Tests and Performance assessments

Rockman claims that tests and surveys "[..] do not demonstrate how well a student has actually learned to navigate through a search strategy process to find, evaluate, use and apply information to meet a specific need" (Rockman 2002, p. 193). Her objections to the use of "objective tests" for the assessment of information literacy learning outcomes are in line with the conclusions of educational reformers who observed that the increasing use of easy to administrate multiple choice and other computerised tests, have led to decreased attention to the learning and assessment of more complex, constituent skills (see for instance Frederiksen 1984, p. 195). In a more recent article on information literacy assessment Megan Oakleaf affirms this disadvantage of "Fixed-Choice Tests" (Oakleaf 2008). Oakleaf explains how a testing culture has the theoretical background of scientific measurement in which "learning tasks should be broken down into fundamental building blocks, which instructors would teach, students would learn and instructors would measure" (Oakleaf 2008, p. 234). She notes that tests have some frequently mentioned benefits (easy to score, easy to compare the results of different groups, rather high reliability) but that there are also some real limitations to the use of tests. She identifies the following major disadvantages of the use of tests for information literacy assessment (Oakleaf 2008, pp. 237-238):

- Tests are mostly focused on individual parts of a concept and not on the complete complex construct. Since information literacy is a complex constituent skill (see also Brand Gruwel et al. 2005, p. 488) there is a real danger that tests under-represent the complete construct.
- Tests create "an artificial situation that does not really test how the learner would react in a real-world situation". There is, in other words, a lack of authenticity. Tests tend to over-assess 'knowing what' and under-assess 'knowing how' and create a situation in which students decide to 'learn for the test'.

Although Oakleaf does not discuss the use of tutorial quizzes and questionnaires as assessment tools for information literacy, it seems clear that these disadvantages are also true for these assessment tools. Lorrie Knight (2006) also recognises the shortcomings of the use of tests for the assessment of complex skills. "A recent trend" she writes "is a movement toward authentic assessment, a process that measures how students apply their knowledge to real-time tasks. [..] Authentic assessment is a promising method for the evaluation of information literacy learning outcomes, as it measures not only what students learn through library instruction, but also how the learning is subsequently incorporated in their academic work" (p. 45). According to Knight the concept of "authentic assessment" refers to the same instruments as the performance-based instruments which are distinguished by Gratch-Lindauer (2003, pp. 29-30) and Oakleaf (2008, p. 240 ff.), namely writing assignments, complex tasks, or performances. As Oakleaf explains (2008, p. 240), these assessments have their roots in the constructivist educational theory which states that they are not just instruments for evaluation but also tools for learning - "students should learn by completing an assessment" (Oakleaf 2008, p. 241). More than tests and quizzes, performance assessments are appropriate instruments for the learning and

evaluation of complex and higher order skills such as information literacy (Oakleaf 2008, p. 242-243; Scharf et al. 2007, p. 462). Another important benefit of the use of performance assessments is their contextualised character which promotes the transfer of the acquired skills to other (real life) situations (Oakleaf 2008, p. 243).

2.2 Portfolios

Both Gratch-Lindauer (2003, p. 30) and Oakleaf (2008, p. 240) emphasise that authentic or performance assessment may be aimed at the performance *process* (through monitoring or direct observation) or at the performance *product* (an essay, presentation etc.). A "training portfolio" (Smith and Tillema 2003, p. 627-628) is an example of an instrument that combines the assessment of student products with a reflection on the performance process. Diller and Phelps (2008) give a description of the use of such a portfolio for information literacy assessment. They use ePortfolios that "allow access to a collection of self-selected student work and self-reflection organized around specific learning goals" not for one specific assignment but "as an integral part of the program to measure progress for each student" (Diller and Phelps 2008, pp. 77-79). Other examples in the literature on the use of training portfolios for information literacy assessment are presented by Fourie and Van Niekerk (1999 and 2001) and by Scharf et al. (2007). Fourie and Van Niekerk emphasise the active participation of students by providing examples of students' academic work that prove that they have achieved the learning goals (1999, p. 335; 2001, p. 110) and stress the importance of reflective activities particularly for the improvement of online research skills (1999, p. 337).

2.3 Scoring Rubrics

Most of the authors who report on the use of performance assessments of student products and / or performance processes, also note that for both assessment practices the use of checklists or scoring rubrics is recommended (Fourie and Van Niekerk 1999, p. 342; Gratch-Lindauer 2003 p. 31; Knight 2006, p. 45 ff.; Oakleaf 2008, p. 244 ff.). A scoring rubric may be defined as "a scoring tool for qualitative rating of authentic or complex student work. It includes criteria for rating important dimensions of performance, as well as standards of attainment for those criteria" (Jonsson and Svingby 2007, p. 131). Rubrics are gaining increasing recognition in Higher Education as instruments for objective and authentic assessment of the way in which students "apply their knowledge to real time tasks" (Knight 2006, p. 45). According to the *educational* academic literature these kinds of standard-based assessment tools are particularly useful for the evaluation of general skills that are needed during the work on open-ended study tasks, which means that in some way students influence the learning route they use to achieve their learning goals (see for instance Elshout-Mohr et al. 2002, pp. 374-375 and Jonsson and Svingby 2007 p. 131). Information problem solving tasks are examples of open-ended study tasks because students often have to choose their own research topics, or at least formulate their own focus on the topic that is given by their professor. Mertler (2001) claims that there are two types of rubrics: holistic and analytic. A holistic rubric is used for grading the overall process or product, without judging the component parts ('dimensions' or 'traits') separately. He contrasts this with an analytic rubric where "the teacher scores separate individual parts of the product or performance first, then sums the individual scores to obtain a total score". The use of a holistic rubric is of course less time consuming but an analytic rubric has the benefit of giving more detailed feedback. In educational practice grading is sometimes restricted to the use of one specific trait of an analytic scoring rubric (Nitko and Brookhart 2007, p. 269).

The information science literature on the use of information literacy scoring rubrics mentions the following advantages:

- Reduced subjectivity with the grading of student assignments thanks to detailed descriptions of the levels of attainment (Oakleaf 2008, pp. 245-246).
- The availability of a scoring rubric makes it easier to give direct and detailed feedback (Oakleaf 2008, p. 245).
- The development of a scoring rubric by teaching staff stimulates the creation of shared information competency beliefs (Knight 2006, p. 52; Oakleaf 2008, p. 246). Scharf et al. (2007, p. 469) notice that reliability may be understood as a consequence of these interactions.
- Distribution of the rubric at the start of the assignment and discussions on it with students, allow them to understand the expectations of their instructors (Scharf et al. 2007, p. 471; Oakleaf 2008, p. 245).
- Students can use the rubrics for self evaluation during their assignments (Oakleaf 2008, p. 246).
- Standardised rubrics make it possible to evaluate student learning across time or multiple programs (Oakleaf 2008, pp. 245-246).

Of course there are also disadvantages. The disadvantage that is most often mentioned in the literature is the fact that the development of the rubric or (in the case of a standardised rubric) learning to work with it is very time consuming (Knight 2006, p. 52; Diller and Phelps 2008, p. 82). In addition, graders must be trained or 'normed' on the rubric before they can work with it and the 'norming' process may be time consuming (Oakleaf 2009, pp. 975-976). On the other hand, these instructions and interactions create awareness of relevant criteria and shared information competency beliefs, as mentioned earlier.

2.4 Questionnaires

In the context of information literacy assessment, questionnaires are often used for the measurement of students' confidence in solving information problems, which is part of the affective dimension of information literacy (Cochrane 2006, pp. 106-111; Kurbanoglu et al. 2006, pp. 731-732; Monoi et al. 2005). As Gross and Latham (2007, p. 349) remark, the data that is collected with these kind of instruments rely heavily on "honesty, openness, and motivation of respondents". This makes questionnaires less usable for summative performance assessment in credit bearing courses. Recently Timmers and Glas (2010) reported the development of a questionnaire in Dutch for the measurement of students' information-seeking behaviour. Their instrument can also be characterised as a set of "self-report scales" that can be used "to compare groups, monitor populations and to determine effects of interventions" (Timmers and Glas 2010, p. 63), but they do not mention examination. For the same reasons it seems plausible that this inappropriateness for credit bearing performance assessment is also true for the use of research diaries, although these can be suitable instruments for the promotion of learning. Finally, the affective dimension of information literacy may also be assessed by the reflective parts of a portfolio (Fourie and Van Niekerk 2001, p. 111).

3. Research approach

The main purpose of the research project was the development of a tested assessment tool for the grading of the students' performance in information problem solving during their study tasks. The literature review in the previous section suggests that scoring rubrics are good grading tools for this. I preferred to construct an analytic scoring rubric because it can provide extensive feedback to the students. The instrument might also be useful for the evaluation of the reflective part of a course related portfolio, but because the use of portfolios is not commonplace in Dutch Higher Education I limited the assessment tool to the evaluation of student products and reported search strategies.

3.1 Construction of the scoring rubric

Craig Mertler (2001) gives a step-by-step procedure for the construction of an analytic scoring rubric. However, I could not employ his design model in its entirety because his model was meant for the construction of scoring rubrics for concrete and specific assignments, while the rubric that I wanted to design was supposed to be more generally applicable. For that reason I could not begin with what Mertler calls the "re-examination" of the learning objectives where a review of these objectives is needed before one undertakes the construction of the rubric (step 1). Instead, I started with the identification (Mertlers step 2) and characterisation (Mertlers step 3) of the attributes that I wanted to use for grading (the 'criteria'). According to Mertlers procedures the next steps should be the descriptions of excellent and poor work for each criterion (Mertlers step 4b) and the description of other levels on the continuum (Mertlers step 5b). Before 5b I had to add one extra step, the decision about the number of performance levels to be used. Mertlers step 6 refers to the collection of samples of student work that exemplify each level. Because my rubric was not related to an existing assignment, I could not collect sample work and had to test the first prototype of the scoring rubric with the results of an assignment that was given by two of my colleagues at The Hague University. The revision of the scoring rubric as necessary (Mertlers step 7) was the last step of the design process. Table 1 gives a summary of the design procedure that I used.

Table 1: Step-by-step procedure for the design of the scoring rubric

a. Identification of the criteria to be used. b. Description of 'professional behaviour' for each individual criterion. c. Description of 'insufficient behaviour' for each individual criterion.	Brain-storming
d. Choice of the number of performance levels to be used in the scoring rubric. e. Description of the intermediate levels of behaviour for each individual criterion.	Elaboration
f. Testing of the scoring rubric with examples of student work at different levels. g. Revision of the scoring rubric as necessary.	Testing

3.1.1 Brainstorming phase

In the brainstorming phase three methods were used to identify the major criteria for the scoring rubric and to describe the 'professional' and the 'insufficient' behaviour (steps a, b and c):

- Analysis of the ACRL standards (2000) and their performance indicators.

- Tracing of the search behaviour of students during their course Desk Research (a ten week full time learning unit of the Department of Information Studies, The Hague University of Applied Sciences).
- Review with colleagues from the Department of Information Studies of The Hague University and the same with three colleagues from other departments (European Studies, Business Economics and Nursing).

The starting point to identify the criteria to be used was the analyses of the ACRL Standards and their performance indicators. Instead of drawing from the text of the Standards directly I used the book by Teresa Neely (2006) because of the many examples of educational practice that she describes. These 'good practices' were analysed to formulate the criteria and to describe the professional behaviour for each criterion. Another good source in this phase of the research was the journal article on direct assessment of information literacy from Scharf et al. (2007). Criteria that emerged during the first draft version of the scoring rubric were:

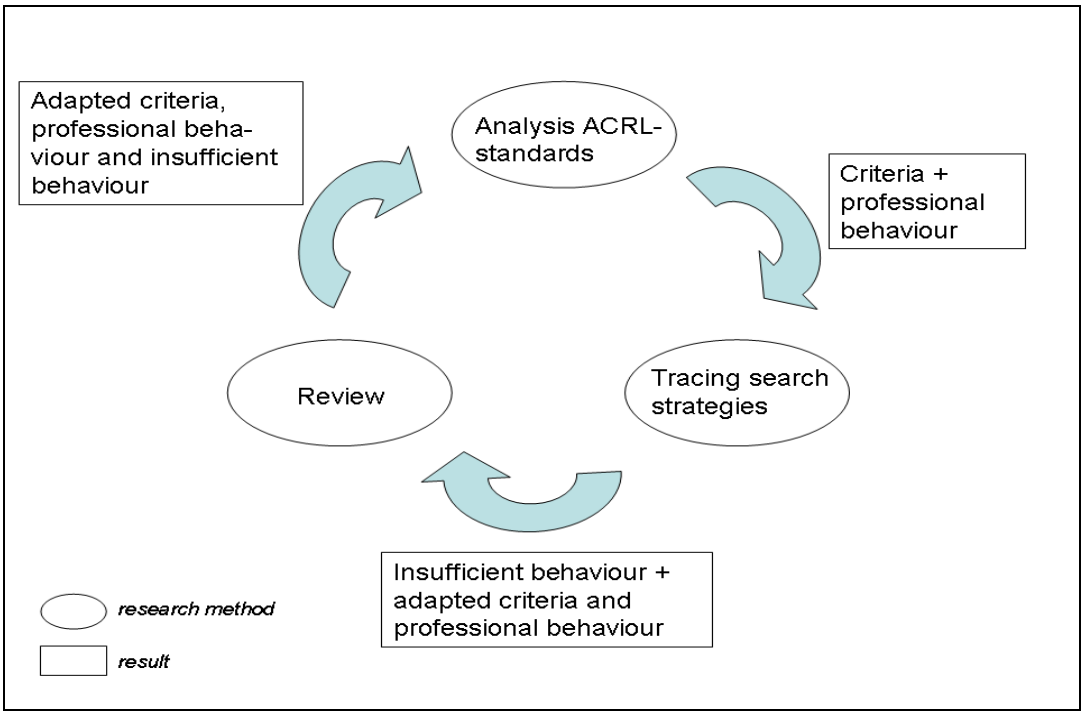
1. Orientation on the research topic
2. Use of desk research as a research method
3. Reference list
4. Quality of primary sources used (books, articles, websites)
5. In text citations
6. Creation of new knowledge
7. Search terms / keywords
8. Secondary resources used (bibliographic tools, web directories, search engines etc.).

Criteria 1-6 can be demonstrated in the student product (the paper, presentation or report), the last two criteria can only be observed when the teachers ask for a separate Search Strategy Report.

Descriptions of 'insufficient behaviour' were initially derived from my own teaching experiences in the previously mentioned ten week Desk Research course. During this course that is offered by my department of The Hague University of Applied Sciences students are asked to formulate and to reformulate the search strategy that they use. The tracing of these search strategies during courses in the academic year 2007-2008 provided me with a lot of examples of insufficient behaviour. The overview with criteria and descriptions of professional and insufficient behaviour was presented to colleagues from my own department as well as to those of three other departments. These colleagues' feedback ensured the validity of the identified scoring criteria.

Research methods (analysis, tracing, reviews) as well as research steps (identification of the criteria and description of the behaviour) were not employed in a straight linear way. The design activities started with the analysis of the ACRL standards but the complete process was repeated three times to improve the draft version of the scoring rubric ('iteration'). The outcomes of the reviews with colleagues for instance led to a reformulation of the criteria and the descriptions of professional behaviour, but they also led to changes in the description of the students' search strategies and the insufficient behaviour.

Figure 1: identification of the criteria and description of professional and insufficient behaviour in the brainstorming phase of the design process



3.1.2 Elaboration phase

To prevent the scoring rubric from becoming too complex to use for end users who are not information professionals themselves (i.e. teachers and students), I decided to restrict it to three levels. Based on the descriptions of the professional and insufficient behaviour, I described the behaviour for the intermediate level for each criterion which was called "moderate". This resulted in a draft version of the scoring rubric (in Dutch) that was tested with two colleagues during a ten week course and a cohort of 15 students.

3.1.3 Testing phase

The research question in the testing phase was:

- How high is consensus between two markers for each criterion if they both grade the same student products with the scoring rubric?

Fifteen students participated in the assignment, which was to write a country profile as preparation for a job or a traineeship in a foreign country. Two teachers coached the writing process of the country reports and at the end they scored the student products with the draft version of the scoring rubric in order to assess the way in which the students had sought information and processed it to solve their information problem (one teacher graded eight country profiles, the other seven). As the researcher and the constructor of the scoring rubric I graded all 15 profiles. The scores of the teacher/coaches were compared with my scores.

The total of the student products for each teacher/grader (7 and 8 respectively) was not high enough for serious statistical analysis of the outcomes of the scoring by the tutors. That is why the answers to the research question were only used as an indication of the need to improve the formulation of the criteria and/or the descriptions of professional, moderate or insufficient behaviour.

The main conclusions after the testing phase were:

- Formulations of the behaviour for some criteria were not clear and should be improved.
- Criteria 1 and 2 were hard to distinguish for markers and can be taken together. Of course, this means that the formulation of the professional, moderate and insufficient behaviour for criterion 1 should also be revised.
- For the markers it was sometimes hard to distinguish between the three levels. They would have preferred more scoring levels. Simply adding one or two levels was not a good solution because insufficient information behaviour can be manifested in many ways. After further consultation of the literature on assessment methods I found an alternative which involved limiting the descriptions to the professional and insufficient behaviour and adding a 6 point Likert scale to do the scoring. Figure 2 gives a detailed example for what finally became the second criterion, 'Reference list'.

The complete scoring rubric is presented as an appendix of this article. For each criterion (columns 1 and 2) the professional and the insufficient behaviour are described. The purpose of these descriptions is to make clear for graders which elements they should pay attention to. Descriptions of the behaviour are accompanied by check boxes to make the grading process easier to carry out. Sometimes the insufficient behaviour is accompanied by two or more check boxes (see for instance the example in figure 2). In these cases insufficient behaviour is demonstrated if one or more of its descriptions are applicable.

After checking the professional and the insufficient behaviour (and all the kinds of behaviour in between) for a criterion, the graders are supposed to score it on the 6 point Likert scale ('very good' to 'very bad'). Finally, they can translate their scoring into a grade from 1-10 or 1-20, depending on the weight that is given to the criterion. Translated to the Dutch educational system the scores from the Likert scale correspond to the grades as follows:
Very good = 10 / 20
Good = 8 / 16
Sufficient = 6 / 12
Poor = 5 / 10
Bad = 3 / 6
Very bad = 1 /2

For the use of the scoring rubric the application of the grades from the last column is not obligatory. Graders can decide to restrict their evaluation of the student performance to the 6 point Likert scale. With these adaptations the prototype of the scoring rubric was finalised and tested in the last stage of the research project.

Figure 2: Layout of the final version of the scoring rubric (translated from the Dutch version)

Criterion		*Professional behaviour*	*Insufficient behaviour*	
2	Reference list	☐ The student product has a reference list that is complete and the citation style is used correctly. With the reference list it is easy to identify the documents that the student used. Remark: the last point is more important than a correct bibliographic description in accordance with a standard citation style. However, for the score 'very good' the citation style must also be used correctly.	☐ There is no reference list in the student product and / or ☐ The reference list is not complete (documents that are cited in the text are not listed in the reference list) or ☐ Important bibliographic data (title, author, year of publication) are missing. An example that often recurs in educational practice: for internet resources only the URL is mentioned.	Grade 1-10=
Score:		0 very good 0 good 0 sufficient	0 poor 0 bad 0 very bad	

3.2 Evaluation of the prototype

The evaluation of the scoring rubric concentrated on three facets:

- Usefulness in educational practice.
- Efficiency in use which was interpreted as the time that markers need to make up the gradings.
- Interrater reliability, which refers to the “level of agreement between a particular set of judges on a particular instrument at a particular time” (Stemler 2004).

The final version of the prototype was employed during winter and spring 2009 in undergraduate courses at two different universities. At the Saxion School of Marketing & International Management (Enschede) the scoring rubric was evaluated by eight teachers who had a background in marketing and/or management, but were not familiar with information literacy. At the start of the project an educational researcher, who is doing a PhD on information literacy, explained the meaning and the scope of each criterion to ensure that the markers would share a common conception of the phenomenon information problem solving. After having graded five student papers with the scoring rubric, the teachers answered a short questionnaire on the usefulness and efficiency of the rubric. The main findings were as follows:

- Seven of the eight teachers were of the opinion that the criteria of the scoring rubric were relevant for grading information literacy and also that no criteria were missing.
- Seven of the eight teachers indicated that the scoring rubric helped them to focus on the relevant aspects for grading information literacy.
- The time needed to score a student product with the rubric was divided into four options. Six respondents selected the first option, 0-5 minutes, and two respondents selected the second option 5-10 minutes. By contrast, no respondent selected the remaining options, 10-15 minutes and greater than 15 minutes, but the reason for this could not be examined further because the teachers completed the questionnaire anonymously.

Six of the eight teachers indicated that they were willing to use the scoring rubric in the future for instruction or formative (diagnostic) student measurement. Most of them were not sure if they were going to use it for examination (summative assessment). Unfortunately none of the respondents elaborated on this point. One of the Saxion respondents commented that she/he would have appreciated it if the behaviour for the intermediate levels had been described. I shall discuss this point in the ‘Conclusions and discussion’ section of this article.

In the Faculty of Military Sciences at the Netherlands Defence Academy (NLDA, Breda) the scoring rubric was used by two professors in information science who both graded the same set of 27 literature reviews. They used the 6 point Likert scale as well as the grading column to give a final grading (1-10). Both professors were quite familiar with the information literacy concept and therefore I decided not to organise a session with them to explain the relevance and the meanings of the various criteria. However, I did send them an instruction on the use of the scoring rubric. The work on the assignments by the students was coached by the professors over a period of seventeen weeks.

The main research question for this field test was the level of "interrater reliability" between the two markers in grading the literature reviews. In statistical theory it is quite normal to distinguish consensus in grading (which means 'exact agreement') from consistency (Stemler 2004). Consistency in grading means that the relative standing of the student products for two or more graders are highly correlated. High graded products by grader 1 are, in other words, also high graded by grader 2, although the two graders do not have to agree on the absolute grading.

The assumption at the beginning of the field test was that consensus and consistency on the grading by the two professors would be appropriate because of their familiarity with the information literacy concept. The results of the grading process were copied and sent to the researcher / constructor of the scoring rubric and were processed using Excel. The analysis of the data was done by generating overviews in crosstabs for each criterion and for the final grading. The main finding of this analysis was that the two markers did not reach a high level of "absolute agreement" but that "adjacent agreement" (which means that the markers did not differ more than 1 point on the 6 point Likert scale or the 10 point grading scale) was quite acceptable. Steven Stemler (2004) reports that adjacent agreement is commonly used for measuring consensus estimates. In the context of the 'norming' process of markers or judges, he argues that it is often too time consuming to train them to the point of exact agreement.

In the NLDA case adjacent agreement was acceptable for most criteria (85% or more) as well as for the final grading (>80%). Only for criterion 7 (the use of secondary sources; search tools) was the adjacent agreement much lower, owing to the fact that the use of some secondary resources was a requirement in the assignment. For that reason the interpretation of one marker was that all students performed quite well for that criterion, while the other marker paid more attention to the use of search tools and bibliographic instruments that were not given in the assignment. The exact values for absolute and adjacent agreement are presented in columns 3 and 4 of table 2.

Percentages of agreement were not only calculated for the exact scores but also for the pass-fail decisions to establish "decision consistency" (Nitko and Brookhart 2007, p. 79). For summative assessments that function as examinations, agreement on the pass-fail decisions is highly relevant because of the consequences for the students' study progress. In this case the decision point for the Likert scale that has been used was sufficient-poor, but the decision point might vary from assignment to assignment. For first year students, for instance, a 'poor' score on the Likert scale might perhaps be good enough to pass.

For three criteria the agreement on the pass-fail decision (table 2, column 5) was lower than 85% and this was deemed problematic. The evaluation of this finding with the participating professors pointed to the fact that the markers had overlooked some mistakes during their grading, something which can easily happen for criteria 2 and 4 (reference list and in text citations). Criterion 5 (creation of new knowledge) is in itself influenced by subjective interpretation. These findings led to two main conclusions:

- For summative (credit bearing) assessments it is strongly recommended to have the grading done by at least two markers and to compare the scores before determining the final grading. Due to

time constraints this is unfortunately not common practice for all assignments/assessments in Dutch Higher Education.

- For summative (credit bearing) assessments it is also recommended to use the scoring rubric as a whole, rather than employing the isolated criteria separately.

Final gradings based on all seven criteria appear to be more reliable than gradings based on isolated criteria. This is confirmed by the fact that the final gradings for the two graders were more consistent than the scores for the separated criteria. Consistency for the final gradings was estimated by calculating Cronbach's alpha (Stemler 2004). The resulting score 0.78 may be considered as rather good. The relatively high consistency of the final gradings is without doubt caused by the compensation of possible grading faults when all seven criteria are used.

Table 2: consensus estimates between grader 1 and grader 2 for the NLDA case

	n =	Absolute agreement	Adjacent agreement	Agreement on pass-fail ("Decision consistency")
Criterion 1	27	11 (41%)	24 (89%)	23 (85%)
Criterion 2	27	7 (26%)	23 (85%)	18 (67%)
Criterion 3	27	11 (41%)	23 (85%)	27 (100%)
Criterion 4	27	12 (44%)	23 (85%)	20 (74%)
Criterion 5	27	5 (19%)	23 (85%)	19 (70%)
Criterion 6*	25	14 (56%)	24 (96%)	23 (92%)
Criterion 7*	25	4 (16%)	17 (68%)	23 (92%)
Final grading (1-10)	27	3 (11%)	22 (81%)	25 (93%)

* Criteria 6 and 7 could only be scored for 25 students because 2 students did not report their search strategy.

At the end of the NLDA field test I discussed the results with the two professors in information science. The main conclusions drawn from this evaluation were that the seven criteria were well chosen and that they are enough for grading information literacy competency. They did miss, however, criteria for the grading of the more formal aspects of the reviews (layout, spelling and grammar) which are indeed not a part of the information literacy conception that I used in this research. The descriptions of behaviour on two levels were in their opinion very helpful when using the scoring scheme.

4. Limitations of the research

The research focused on the construction of an *assessment tool* for information literacy behaviour. Such an assessment instrument can be a good starting point for the design of information literacy training programmes or for integrating information literacy in discipline based curricula. However, the availability of the scoring rubric is not enough for good information literacy education. For that purpose there is still a lot of work to be done on development, evaluation and probably the redesigning of existing curricula and training programmes.

The scoring rubric is restricted to the assessment of the information literacy facet of student products. Other facets like professional knowledge and other general skills, such as writing skills are not incorporated in the grading tool. Overall academic staff might recognise the significance of assessing information problem solving skills of their students, but they also need to assess the subject knowledge from their discipline which is not addressed by the information literacy scoring rubric that I constructed.

A theoretical weakness of my scoring rubric is the restriction to a task oriented conception of information literacy or, in the words of Christine Bruce, the "Information Process Conception" (Bruce 1998, pp. 32-33). Another conception of information literacy that is often used in the research literature is the Knowledge Base Conception (see for instance Maybee 2006, p. 83) which includes Bruce's Knowledge Construction Conception and Knowledge Extension Conception (Bruce 1998, pp. 35-37). The knowledge base conception as it is described by Maybee in his study on the information use conceptions of undergraduate students, refers to a long-term conception of information literacy which includes the building of personal meanings and knowledge that can be transferred to other situations (2006, p. 83). For the assessment of the knowledge base conception of information literacy another (more longitudinal) instrument would be needed, for instance using portfolio assessment techniques with more emphasis on reflective activities by the students and emphasis on their personal development.

The evaluation of the scoring rubric was limited to the teachers' experiences (usefulness for educational practice, efficiency in use and interrater reliability). It would be useful to investigate the extent to which students appreciate the instrument for instruction, feedback and self assessment in a follow-up study. Finally, it should be noted that the number of participating students and teachers for the Saxion field tests was relatively low. This precludes any generalisation of the outcomes of the Saxion evaluation to other educational situations. The number of student products that were graded by the NLDA professors was just high enough to calculate statistical values for interrater consensus and consistency. This was confirmed in a personal communication with Dr. Ron Oostdam from the SCO-Kohnstamm Institute (University of Amsterdam).

5. Conclusions and discussion

Scholarly literature on information literacy performance assessment suggests that analytical scoring rubrics are good tools for the grading of the students' information behaviour during their study tasks. In this research I developed such an assessment tool and evaluated it from a teachers' viewpoint. Teachers from the Saxion School of Marketing & International Management as well as the information science professors from the Netherlands Defence Academy (NLDA) affirmed the content validity of the seven criteria and claimed that the scoring scheme was helpful to keep them focused on the assessment of information literacy competences.

One of the Saxion respondents commented that she/he would have liked to have seen the behaviour for the intermediate levels described in the rubric. As mentioned in section 3.1.3 I limited the descriptions to the highest and to the lowest levels and added a Likert scale to do the scoring because insufficient behaviour can be manifested in many ways. This may be due to the fact that the selected criteria are still rather generic and that very often different measures for insufficient behaviour, e.g. the 'check boxes', are used. Splitting the criteria into more detailed criteria would make the rubric, however, much more extensive. For example criterion 3 (quality of the primary sources) should have been worked out to at least five separated criteria. This could deter teachers from using the rubric as confirmed by the feedback from the teachers who tested it.

Reliability of the scoring rubric depends on the conception of interrater reliability that is used**.** The two graders from the NLDA did not reach high degrees of absolute consensus on the 6 point Likert scale when grading the same set of 27 student products, although the degree of 'adjacent agreement' was quite acceptable for almost all criteria as well as for the final grading. The agreements on the pass-fail decisions were acceptable for the final grading but problematic for three criteria. This led me to the conclusion that for credit bearing assessments it is recommended to have the grading done by at least two different markers and to discuss the results before determining the final gradings, although this is not common practice in Dutch Higher Education. A second recommendation generated by this

research is the use of the complete set of seven criteria for credit bearing decisions because the consistency between the two graders was much better for the final gradings than for the separate criteria.

Acknowledgements

The author is grateful to his colleagues from different universities who participated in the reviews and the tests of the rubric. In particular, he would like to thank Caroline Timmers from Saxion and Peter Jongejan and Maarten van Veen from the Netherlands Defence Academy. Without their contributions it would have been almost impossible to complete this research. My thanks go to Ron Oostdam from the SCO-Kohnstamm Institute (University of Amsterdam) for his tips on solving educational research problems and Albert Boekhorst (University of Amsterdam) for his comments during the research process.

References

ACRL 2000. *Information literacy competency standards for Higher Education*. Chicago: American Library Association.

Bloom, B. (Ed.) 1956. *Taxonomy of educational objectives : the classification of educational goals. Handbook I, cognitive domain*. London: Longman.

Boekhorst, A. 2003. Becoming information literate in The Netherlands. *Library Review* 52(7), pp. 298-309.

Brand-Gruwel, S. et al., 2005. Information problem solving by experts and novices: analysis of a complex cognitive skill. *Computers in Human Behavior* 21(3), pp. 487-508.

Bruce, C. 1998. The phenomenon of information literacy. *Higher Education Research & Development* 17(1), pp. 25-43.

Cochrane, C. 2006. Embedding information literacy in an undergraduate management degree: lecturers' and students' perspectives. *Education for Information* 24(2), pp. 97-123.

Diller, K. and Phelps, S. 2008. Learning outcomes, portfolios, and rubrics, oh my! Authentic assessment of an information literacy program. *Portal: Libraries and the Academy* 8(1), pp. 75-89.

Dirkx, A. et al., 2006. In drie uur bewust onbekwaam! *InformatieProfessional* 10(11), pp. 16-19.

Dochy, F. and McDowell, L. 1997. Assessment as a tool for learning. *Studies in Educational Evaluation* 23(4), pp. 279-298.

Dochy, F. et al., 2005. Students' perceptions of a problem-based learning environment. *Learning Environments Research* 8(1), pp. 41-66.

Drent, M., and Timmers, C. 2006. Informatieproblemen oplossen: een digitale en leerzame toets. *OCN Online Conferentie Nederland, 2006*.

Driessen, E. and Vleuten, C. van der. 2000. Matching student assessment to problem-based learning: lessons from experience in a law faculty. *Studies in Continuing Education* 22(2), pp. 235-248.

Elshout-Mohr, M. et al., 2002. Student assessment within the context of constructivist educational settings. *Studies in Educational Evaluation* 28(4), pp. 369-390.

Fourie, I. and Niekerk, D. van. 1999. Using portfolio assessment in a module in research information skills. *Education for Information* 17(4), pp. 333-352.

Fourie, I. and Niekerk, D. van. 2001. Follow-up on the use of portfolio assessment for a module in research information skills: an analysis of its value. *Education for Information* 19(2), pp. 107-126.

Frederiksen, N. 1984. The real test bias: influences of testing on teaching and learning. *American Psychologist* 39(3), pp. 193-202.

Gratch-Lindauer, B. 2003. Selecting and developing assessment tools. In: Avery, E. (Ed.) *Assessing student learning outcomes for information literacy instruction in academic institutions*. Chicago: American Library Association, pp. 22-39.

Gross, M. and Latham, D. 2007. Attaining information literacy: an investigation of the relationship between skill level, self-estimates of skill, and library anxiety. *Library and Information Science Research* 29(3), pp. 332-353.

Jonsson, A. and Svingby, G. 2007. The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review* 2(2), pp. 130-144.

Knight, L. 2006. Using rubrics to assess information literacy. *Reference Services Review* 34(1), pp. 43-55.

Kurbanoglu, S. et al., 2006. Developing the information literacy self-efficacy scale. *Journal of Documentation* 62(6), pp. 730-743.

Maybee, C. 2006. Undergraduate perceptions of information use: the basis for creating user-centered student information literacy instruction. *Journal of Academic Librarianship* 32(1), pp. 79-85.

Mertler, C. 2001. Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation* 7(25). [Online] Available at: http://pareonline.net/getvn.asp?v=7&n=25 [Accessed: 1 October 2009].

Monoi, S. et al., 2005. Online searching skills: development of an inventory to assess self-efficacy. *The Journal of Academic Librarianship* 31(2), pp. 98-105.

Neely, T. 2006. *Information literacy assessment: standards-based tools and assignments*. Chicago: American Library Association.

Nitko, A. and Brookhart, S. 2007. *Educational assessment of students*. 5th ed. Upper Saddle River, N.J.: Pearson Education.

Oakleaf, M. 2008. Dangers and opportunities: a conceptual map of information literacy assessment approaches. *Portal: Libraries and the Academy* 8(3), pp. 233-253.

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.

Rockman, I. 2002. Strengthening connections between information literacy, general education and assessment efforts. *Library trends* 51(2), pp. 185-198.

Scharf, D. et al., 2007. Direct assessment of information literacy using writing portfolios. *Journal of Academic Librarianship* 33(4), pp. 462-477.

Smith, K. and Tillema, H. 2003. Clarifying different types of portfolio use. *Assessment & Evaluation in Higher Education* 28(6), pp. 625-648.

Stemler, S. 2004. A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation* 9(4). [Online] Available at: http://www.pareonline.net/getvn.asp?v=9&n=4 [Accessed: 1 October 2009].

Stubbings, R. and Franklin, G. 2006. Does advocacy help to embed information literacy into the curriculum? A case study. *Italics* 5(1). [Online]. Available at:

http://www.ics.heacademy.ac.uk/italics/vol5-1/pdf/stubbings-franklin-final.pdf [Accessed: 25 February 2010].

Tigelaar, D. et al., 2004. The development and validation of a framework for teaching competencies in Higher Education. *Higher Education* 48(2), pp. 253-268.

Timmers, C. and Glas, C. 2010. Developing scales for information-seeking behaviour. *Journal of Documentation* 66(1), pp. 46-69.

Walraven, A. et al., 2008. Information-problem solving: a review of problems students encounter and instructional solutions. *Computers in Human Behavior* 24(3), pp. 623-648.

Appendix

Scoring rubric for Information Literacy student product **Name teacher / grader:** **Name/ID-No. student:**

Criterion		*Professional behaviour*	*Insufficient behaviour*	
1	Orientation	☐ The student product makes clear that the student did a good orientation on the topic and that he/she formulated his/her own focus on the topic or research question. This is also expressed by the fact that the student formulated one or more good research questions.	☐ The student product makes clear that the student used the question as it was originally formulated in the assignment or student task. The student him/herself did not further explore the question as such. An example of this behaviour is that the student did not define the core key terms and that these terms are supposed to be clear while they are at least multi interpretable.	Grade 1-20=
Score:		0 very good 0 good 0 sufficient	0 poor 0 bad 0 very bad	

Criterion		*Professional behaviour*	*Insufficient behaviour*	
2	Reference list	☐ The student product has a reference list that is complete and the citation style is used correctly. With the reference list it is easy to identify the documents that the student used. Remark: the last point is more important than a correct bibliographic description in accordance with a standard citation style. However, for the score 'very good' the citation style must also be used correctly.	☐ There is no reference list in the student product and / or ☐ The reference list is not complete (documents that are cited in the text are not listed in the reference list) or ☐ Important bibliographic data (title, author, year of publication) are missing. An example that often recurs in educational practice: for internet resources only the URL is mentioned.	Grade 1-10=
Score:		0 very good 0 good 0 sufficient	0 poor 0 bad 0 very bad	

Criterion		*Professional behaviour*	*Insufficient behaviour*	
3	Quality of the primary sources (books, journal articles, websites etc.)	☐ The reference list of the student product makes clear that the student has used relevant, reliable (preferably authentic) and up-to- date information sources that discuss the topic or the question from different points of view.	☐ The information sources the student has used are insignificant, outdated or not relevant enough. An example of 'insignificance' is that the student only used Internet-sites as an information source. And / or … ☐ The information sources the student used are one-sided (too much from one point of view). The student has, for instance, only used government information(.gov-sites) or publications from one particular author.	Grade 1-20=
Score:		0 very good 0 good 0 sufficient	0 poor 0 bad 0 very bad	

Scoring rubric for Information Literacy **Name teacher / grader:** **Name/ID-No. student:**

	Criterion	*Professional behaviour*	*Insufficient behaviour*	
4	In text-citations	In the text of the product it is made clear what information sources the student has used. In the case of a digital student product this is also true for images and audiovisual information.	The student has used someone else's work (text fragments, images, audiovisuals) in his / her own product without reference to the original source. Even if this was done unintentionally, strictly speaking this is plagiarism.	Grade 1-10=
Score:		0 very good 0 good 0 sufficient	0 poor 0 bad 0 very bad	

	Criterion	*Professional behaviour*	*Insufficient behaviour*	
5	Creation of new knowledge out of relevant information	The student product makes clear that the student analyzed information from different resources and that – based on this analysis – he / she formulated new insights, hypotheses or applications. Scope note: practice shows that students succeed in analysing and comparing several information sources, but are not capable of synthesizing the retrieved data into a new insight, hypothesis or application. If so, this criterion should be graded as "sufficient" or "poor".	In the student product the student did not reproduce the content of the retrieved information correctly or clearly and / or paid no attention whatsoever to the analysis of the information sources found and / or used only one information source without discussing the relevance or the reliability of the content, although there is reason for doubt..	Grade 1-20=
Score:		0 very good 0 good 0 sufficient	0 poor 0 bad 0 very bad	

Scoring rubric for Information Literacy

Search Strategy

	Criterion	*Professional behaviour*	*Insufficient behaviour*	
6	Search terms / keywords	The student used search terms that are relevant for the topic or the research question. He / she used relevant synonyms, search terms in English and from the professional jargon.	The student used search terms that are too general (non professional) and / or the student did not use relevant synonyms, associated terms or search terms in English.	Grade 1-10=
Score:		0 very good 0 good 0 sufficient	0 poor 0 bad 0 very bad	
	Criterion	*Professional behaviour*	*Insufficient behaviour*	
7	Use of secondary sources	The student used a variety of secondary sources (search engines, books for tracking citations, scholarly journals, databases, social networks). If necessary he / she used an interlibrary loan to obtain the materials needed.	The student only used information sources that are easily accessible. For instance: he / she only used • The "quick search"-box of a general search engine and / or • Materials provided by his / her professor.	Grade 1-10=
Score:		0 very good 0 good 0 sufficient	0 poor 0 bad 0 very bad	

Total score (maximum 100) =

Van Helvoort. 2010. *Journal of Information Literacy.* 4(1).
http://ojs.lboro.ac.uk/ojs/index.php/JIL/article/view/PRA-V4-I1-2010-2