Journal of Education and Human Development
March 2016, Vol. 5, No. 1, pp. 86-96
ISSN: 2334-296X (Print), 2334-2978 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/jehd.v5n1a9
URL: http://dx.doi.org/10.15640/jehd.v5n1a9

Self-Assessment Scores Increase in Parallel with Objective Performance Measures: The Case of Information Literacy in the Context of a Randomized Blended Learning Program Evaluation

Nikolas Leichner[1], Anne-Kathrin Mayer[2], Johannes Peter[3] & Günter Krampen[4]

Abstract

In the context of a randomized evaluation study for an information literacy instruction program with *N* = 67 participants, changes in objective information literacy performance measures and self-assessed abilityare examined. Results show that scores on both kinds of instruments increase nearly in parallel. Another finding is that self-assessed abilities and objective measures are moderately correlated before participation in the instruction program. The correlation, however, diminishes after participation in the instruction program; this finding is attributed to the unspecific nature of the scale which seems unsuitable in light of the higher performance levels after participation.

Keywords:self-assessment, information literacy, college students, objective performance measures.

1.Introduction

Self-assessment of ability is an important issue in educational psychology due to its interrelation with actual performance. For example, Chamorro-Premuzic and Furnham (2006) found that self-assessed intelligence scores explained incremental variance components in academic performance beyond objective intelligence scores. Furthermore, there is indication that (positive) self-evaluation can increase performance (Assor & Connell, 1992; Bandura, 2010).

Additionally, self-assessment instruments are sometimes used complementarily or even as an alternative to objective performance measures, when no objective tests are available or their use requires disproportionate resources (Skeff, Stratos, & Bergen, 1992). In the field of information literacy, self-assessment instruments are often used as the single measure (Schilling & Applegate, 2012).

However, the validity of self-assessments with regard to actual performance is questionable. Findings indicate that persons are at best moderately able to estimate their own performance (e.g., the meta-analysis by Falchikov & Boud, 1989). Unfortunately, only a few studies examined the change of subjective performance measures over study time. Exceptions are the studies by Arnold, Willoughby, and Calkins (1985) and Cassidy and Eachus (2000) indicating that scores on self-assessment instruments increase with duration of study time.

[1]Leibniz Institute for Psychology Information, Universitätsring 15, 54296 Trier, Germany.
[2]Leibniz Institute for Psychology Information, Universitätsring 15, 54296 Trier, Germany.
[3]Leibniz Institute for Psychology Information, Universitätsring 15, 54296 Trier, Germany.
[4]Leibniz Institute for Psychology Information, Universitätsring 15, 54296 Trier, Germany and University of Trier, Universitätsring 15, 54296 Trier, Germany.

In light of these findings, aim of this paper is to examine how objective and subjective measures as well as their relationship change by participation in an information literacy instruction program. Information literacy is defined as the ability to recognize an information need and to subsequently locate, evaluate and use the needed information (Association of College and Research Libraries [ACRL], 1989).

For this purpose, both kinds of measures will be assessed at three times of measurement in the course of the instruction program. In addition to examining the effectiveness of the information literacy instruction program, it will be examined, (a) whether scores on both kinds of measures increase in a parallel way, (b) whether increase on both kinds of instruments (in terms of difference scores) is correlated, and (c) to which extent self-assessed abilities and objective measures of information literacy are correlated.

Self-assessment and performance

As mentioned above, self-assessments of abilities are often used in educational psychology for a variety of reasons. Nonetheless, the validity of these measures is often questionable. The accuracy of self-assessment (i.e., the relationship between self-assessed abilities and objective performance measures) is often only moderate, indicating that people have problems estimating their own performance accurately. A closer look, however, reveals enormous variation in the accuracy of self-assessments among studies. For example, a recent study aggregating several meta-analyses found a mean correlation between self-assessed performance and objective performance measures of M = .29, but also substantial variation: coefficients ranged from .09 to .63 (Zell & Krizan, 2014). Several factors influencing self-assessment accuracy were identified. A crucial factor seems to be performance level: higher achieving students (Lew, Alwis, & Schmidt, 2010) and those taking advanced courses (Falchikov & Boud, 1989) are able to assess their performance more accurately. One explanation is that lower achieving individuals lack the metacognitive skills required for adequate self-assessment. Kruger and Dunning (1999) were able to show that metacognitive skills predicted inflated self-assessments even when controlling for actual test performance. Another finding was that improving participants' metacognitive skills also enhances self-assessment accuracy. This implies that metacognitive skills are a critical moderating variable for the relationship between performance level and self-assessment accuracy.

The change of self-assessment scores over time has, to our knowledge, only been the topic of a few studies. Arnold et al. (1985) examined the self-assessment scores of 211 medical students over a period of four years. The authors found that self-assessment scores as well as faculty ratings increased over time, however, the relationship between the two kinds of measures diminished over time. In a study by Cassidy and Eachus (2000), 130 undergraduate students from various disciplines provided a self-assessment of their research methods proficiency before and after completion of a research methods module. It was found that self-assessed proficiency improved markedly by completing the module. Furthermore, self-assessed proficiency was significantly correlated with the module grade. In a study with 108 schoolchildren, it was found that self-assessment scores related to mathematics and reading/English tended to decline between the ages of 7 and 16 (Blatchford, 1997). Besides the considerably younger sample and the longer period of time examined, this study also differs from the other two in the way self-assessment scores were collected. In this study, students just had to indicate whether they considered their own performance to be average, below or above average.

Assessing information literacy

In higher education (e.g. studies of psychology) information literacy, as defined above (ACRL, 1989), implies knowledge of the relevant scholarly information resources (e.g. bibliographic databases) and the ability to use these resources efficiently (ACRL, 2010). Due to a fast growing body of scientific literature, information literacy is considered an important prerequisite for academic achievement (Bowles-Terry, 2013) as well as a fundamental learning outcome of higher education which enables students for lifelong learning (Andretta, 2005). A variety of assessment approaches for information literacy exists. This section aims to provide a short overview of objective assessment approaches and self-assessment measures.

For the objective assessment of information literacy skills, it seems most promising to assess them in a multi-method fashion (cfEid & Diener, 2006) by combining several kinds of assessments. According to McCulley (2009), instruments can either be categorized as standardized tests or performance assessments.

The best-established standardized test format is the multiple choice question (Walsh, 2009). The major advantages of multiple choice tests are their economic use and high reliability (Haladyna& Rodriguez, 2013).

The term performance assessment covers a variety of assessment instruments which require the students to utilize their skills to create a product (McCulley, 2009; Tung, 2010). For the assessment of information literacy, information search tasks have been used (Chen, 2009; Dunn, 2002). The major advantages of performance assessments are that they assess performance in an authentic situation, and they provide the possibility to align measurement instruments with learning goals (Oakleaf, 2008). To evaluate student performance in a standardized way, it is indispensable to create scoring rubrics (Oakleaf, 2008, 2009; Tung, 2010). It may be assumed that multiple choice tests allow the assessment of declarative knowledge, while information search tasks can be used to assess whether this knowledge can be applied in context (procedural knowledge; for a definition, see Anderson, 1996). Information search tasks, additionally, allow the assessment of complex information seeking strategies which cannot be captured by multiple choice items.

Regarding self-assessment of information literacy skills, several instruments exist as well. The most widely used instrument seems to be the information literacy self-efficacy scale developed by Kurbanoglu, Akkoyunlu, and Umay (2006).There are a short and a long version of this scale, containing 17 and 28 items, respectively. The scale assesses self-efficacy beliefs related to several aspects of information literacy, such as the definition of an information need, searching for information, or synthesizing information to create a product. Items require the participants to indicate their agreement to statements like "I feel confident and competent to define the information I need" on a seven point scale. Satisfactory reliability estimates were obtained by the authors.

Another instrument for the self-assessment of information literacy skills is the IL-HUMASS developed by Pinto (2010). The instrument has a similarly broad scope as the scale by Kurbanoglu et al. (2006) and contains 26 items. For each of the activities mentioned (e.g. use of electronic sources), participants are required to indicate their motivation to engage in the relevant activity as well as their self-efficacy related to it.

Current study and hypotheses

The present study on the relations between subjective and objective information literacy assessments is conducted in the context of an information literacy instruction program for psychology students. Thus, Hypothesis 1 refers to the effectiveness of the instruction program: It was expected that the program's participants would subsequently score higher on objective assessments of information literacy than beforehand.

Hypothesis 2 states that self-assessed abilities would increase along with objective performance measures. It was also expected that difference scores on the self-assessment instrument would correlate significantly and in a positive direction with difference scores on the objective performance measures.

Hypothesis 3 refers to the relationship between self-assessed performance and objective measures. As mentioned above, correlations between self-assessments and objective performance measures seem to be moderated by performance level in the sense that correlations increase with performance level. Therefore, it is expected that correlations are higher after participation in the instruction program.

Description of the instruction program

Most university libraries offer information literacy instruction, often in the form of one-shot sessions complemented by individual assistance at the reference desk (Homann, 2003; Mery, Newby, & Peng, 2012). The practice of one-shot sessions has been criticized because a single session is too short to cover information literacy comprehensively (Grassian & Kaplowitz, 2001; Mery et al., 2012). Online courses have been suggested as a viable way to bypass the lack of resources (Mery et al., 2012) because they facilitate teaching large groups of students efficiently (Orr, Williams, & Pennington, 2009), are flexible with regard to time and place (Romero & Barberà, 2011),and seem especially suited for self-directed learning (Moore & Kearsley, 2012). However, online courses are often fraught with high dropout rates (Levy, 2007; Nistor &Neubauer, 2010).

Therefore, combining online and classroom teaching ("Blended Learning", cf Garrison &Kanuka, 2004) is seen as a promising way to overcome problems associated with online teaching alone. There is meta-analytic indication that online teaching is more effective than classroom teaching alone (Bernard, Borokhovski, Schmid, Tamim, & Abrami, 2014).

Based on these findings, the information literacy instruction program evaluated in the present study combined online and classroom teaching (Blended Learning, cf Garrison & Kanuka, 2004). Most of the content was imparted via online materials, while there were two classroom seminars (taking around 90 minutes each) which were designed to give the participants an opportunity to reflect the materials. Another important component of the classroom seminars was the completion of practice information search tasks under the instructors' guidance. The online materials were divided into three modules. The first classroom seminar was scheduled after completion of the first two online modules; the second classroom seminar was scheduled after completion of the third online module. Each classroom seminar was intended to give the participants the opportunity of reflecting and practicingthe content of the previous online modules. The whole course was completed within two weeks.

When designing information literacy instruction, it is generally recommended to use a discipline-specific approach as every discipline or scientific field has its own information resources and publication structure. Knowledge about these "information cultures" as well as domain-specific knowledge is required to search and evaluate information adequately (Grafstein, 2002). The instruction program presented here was designed in a psychology-specific way.

Content of the instruction program was based on the psychology-specific information literacy standards (ACRL, 2010). Some content was added based on our own considerations to make sure the instruction program covers the situation in Germany adequately. Site-specific information (e.g. information about the local library catalogue) was also included. The introductory materials covered scholarly communication patterns in psychology as well as the most relevant document types, e.g. journal articles, edited books, and monographs. A large section of the materials was devoted to a description of the most important information resources for psychologists (particularly the bibliographic databases PsycINFO and PSYNDEX, and web resources like Google Scholar), and the advanced use of these resources (e.g. use of Boolean operators or the thesaurus). Other content dealt with the acquisition of literature (e.g. interlibrary loan, or the use of electronic journal subscriptions) and with criteria for the evaluation of information (e.g. Journal Impact Factor).

Methods

The sample consisted of N = 67 undergraduate psychology students from the University of Trier, Germany, including n = 34 first year, and n = 33 second year students. On average, the participants were M = 21.67 years old (SD = 2.38); 52 of them were female (77%). It was not possible to collect data on course grades.

Instruments

As, to our knowledge, no validated measurement instruments for information literacy were available in German language, new assessment tools were devised. Following our reflections about the assessment of information literacy outlined above, three instruments were developed: a multiple choice test, a set of standardized information search tasks, and a self-assessment scale.

The multiple choice test is based on previous work of our research group (Leichner, Peter, Mayer, & Krampen, 2013). It consists of 35 items with three response options each. When developing the items, we referred to the information literacy standards for psychology which were designed with the facilitation of information literacy assessment in mind (ACRL, 2010). Items deal with the ways of finding scholarly psychologic information (e.g. use of scholarly databases like PsycINFO), the acquisition of literature (e.g. interlibrary loan), and criteria for the evaluation of information (e.g. citation indices). A sample item is provided below.

Which differences exist between Internet search engines (e.g. Google Scholar) and bibliographic databases?

o bibliographic databases usually have a thesaurus search
o Boolean operators can only be used with bibliographic databases
o the order of items on the results page is not affected by the number of clicks on each item

(note: response options 1 and 3 are correct)

The information search tasks were devised by Leichner, Peter, Mayer, and Krampen (2014). The first step in the creation of information search tasks was to establish a task taxonomy describing three task types of ascending difficulty. This was accomplished by designing tasks that differ in the competencies and abilities required to solve them. The concept of additive competencies was used: to successfully complete type 1 tasks certain abilities were required, whereas these and additional competencies were necessary to successfully complete type 2 tasks.

Successful completion of type 3 tasks demands, in turn, of the preceding abilities and additional ones. Furthermore, a questionnaire was created regarding the procedure applied when completing the search tasks (e.g. search engines and search terms used). Finally, two scoring rubrics were developed. According to the search task outcome rubric, scores are awarded if the publications found meet the criteria outlined in the search task (e.g. scores are awarded if the publications found deal with the relevant topic or include a specific age group of participants). According to the search task procedure rubric, scores are awarded for approaching the tasks in an efficient way (e.g. use of relevant filter functions). A detailed description of the taxonomy as well as the questionnaire and the rubrics mentioned above can be found in Leichner et al. (2014). The task taxonomy was used to derive three tasks of each type; for illustrating, a sample type 2 task (medium difficulty) is: *'Are there meta-analyses published after 2005 investigating "risk factors" for the development of a "Posttraumatic stress disorder"? If possible, indicate two publications.'*

For self-assessment of information literacy, the existing scales by Pinto (2010) and Kurbanoglu et al. (2006) seemed unsuitable as they refer to a broad definition of information literacy and include many aspects (e.g. referring to the preparation of a paper) that are neither related to our objective information literacy measures nor to the content of our instruction program. Therefore, a new scale consisting of 10 items was created referring to self-assessed abilities to search and evaluate scholarly information. To ensure that the students would be able to understand the items without further instruction, wording of the items was quite general, so the items capture self-assessed information literacy in a relatively broad sense (see appendix for a translation of all items). Participants were required to answer the items on a five point Likert scale; a "don't know" was also available but rarely used by participants (nearly 80% of the participants did not use this option; no participant used this option more than three times). In a previous study with $N = 173$ psychology students, satisfactoryreliability estimates ($\alpha = .77$; Guttman's $\lambda_6 = .78$) were found.

Procedure

As part of the experimental study, participants were randomly assigned to one of two groups. The experimental group (group 1) consisted of $n = 37$ participants, while the wait list control group (group 2) consisted of $n = 30$ students. The experimental group participated in the two week instruction program right at the beginning of the study. Group 2 followed with a time lag of two weeks; so, the complete study took four weeks. Data were collected three times during the study: At the beginning (t1), after two weeks (t2), and at the end (t3). Data collection was carried out in the computer lab of the University of Trier where the participants completed the instruments via online survey software. At each time of measurement, participants completed three information search tasks (one task of each type ordered by task difficulty), the multiple choice test, and the self-assessment scale. Instruments had to be completed in the order described above (two questionnaires referring to epistemological beliefs that are not part of this paper were completed in between the search tasks and the multiple choice test); at last the self-assessment scale was completed following the recommendations by Rosman, Mayer, and Krampen (2014). After the completion of each information search task, the participants were required to (a) indicate the publications which they had found, and to (b) complete the questionnaire referring to the procedure applied when searching for information. This data later was used to score the search procedure.

Results

Before examining the hypotheses, details referring to the scoring of the instruments and psychometric analyses are reported. The multiple choice test items were scored by giving partial credit if only some of the response options were correctly marked or left out. The total score was computed by averaging the single item scores. The total score was scaled in a way that scores range from 0 to 1. The self-assessment scale was created by averaging the responses to the ten items. As five point Likert scales were used, the total score ranges from 1 to 5.Reliability estimates for both instruments at the three measuring times can be found in Table 1. In addition to the conventional reliability estimates Cronbach's Alpha and split-half reliability, Guttman's Lambda 6 (λ_6) is reported as this estimate is more suitable for less homogenous item sets (Revelle, 2015, p. 230).

Table 1: ***Reliability estimates for the multiple choice test and the self-assessment scale for the three times of measurement***

	t1	t2	t3
Multiple-Choice test			
Cronbach's α	.63	.80	.55
Guttman's Lambda 6	.84	.91	.81
Split-Half (Spearman-Brown)	.64	.78	.57
Self-assessment scale			
Cronbach's α	.84	.79	.67
Guttman's Lambda 6	.86	.81	.73
Split-Half (Spearman-Brown)	.84	.77	.68

The information search tasks were scored by two independent raters. Interrater-reliability as assessed by correlations between the scores awarded by the two raters ranged from r = .62 to r = .87 for the outcome scores, and from r = .72 to r = .92 for the procedure scores. As the variables were of metrical nature, computing conventional measures for interrater agreement (e.g., Kappa) was not possible. For the following analyses, outcome and procedure scores were added up separately for each time of measurement. So, for further analyses, one score for search task outcome and one for search task procedure were available for each time of measurement. These scores were subsequently scaled to a range from 0 to 1.

To determine retest reliability, group 2 scores from t1 and t2 were correlated, as this group participated in the instruction program later. Retest reliability for search task outcome and procedure were r = .53 and r = .63, respectively. Retest reliability for the multiple choice test, and the self-assessment scale were r = .71 and r = .79, respectively.

Mean scores for the four variables are displayed in Table 2.

Table 2: ***Mean scores and (in brackets) standard deviations for the instruments arranged by group and time of measurement***

	t1		t2		t3	
	group 1	group 2	group 1	group 2	group 1	group 2
knowledge test	0.60(0.07)	0.61(0.07)	0.77(0.06)	0.62(0.06)	0.76(0.06)	0.75(0.05)
search task outcome	0.46(0.20)	0.52(0.17)	0.64(0.20)	0.50(0.13)	0.78(0.18)	0.77(0.18)
search task procedure	0.43(0.13)	0.47(0.13)	0.81(0.11)	0.55(0.13)	0.78(0.08)	0.71(0.12)
self-assessment	2.68(0.75)	2.65(0.74)	3.67(0.42)	2.80(0.60)	3.60(0.36)	3.52(0.54)

Hypothesis 1

Before Hypothesis 1 (effectiveness of the instruction program) was examined, group differences before participation in the instruction program were assessed. It was found that the two groups did not differ on the search task outcome scores (t[65] = 1.34, *ns*), the search task procedure scores (t[65] = 1.23, *ns*), or the multiple choice test (t[65] = 0.78, *ns*). To examine Hypothesis 1, separate repeated measures ANOVAs were computed for each information literacy performance indicator: Time of measurement served as within-subjects factor, while group membership served as between-subjects factor. Analysis of the multiple choice test scores revealed a significant interaction of the two factors (F[2,130] = 73.13, $p < 0.01$, η^2 = .53). Repeated contrasts computed separately for each group revealed significant differences between t1 and t2, as well as between t2 and t3 for group 1. For group 2, only the difference between t2 and t3 reached significance level of p< .05.An interaction of the two factors was also found when analyzing search task outcome (F[2,130] = 5.45, $p < .01$, η^2 = .08) and procedure scores (F[2,130] = 37.38, $p < .01$, η^2 = .37).

Repeated contrasts computed separately for each group showed significant increases in outcome scores:group 1 scores increased significantly from t1 to t2 and from t2 to t3 while group 2 scores increased significantly only from t2 to t3. For the procedure scores it was found that group 1 scores increased significantly between t1 and t2; for group 2 scores, both contrasts reached significance ($p < .05$).

Hypothesis 2

To examine Hypothesis 2, the same procedure was repeated for the self-assessment scale. An initial check indicated that the t1 scores of the two groups did not differ ($t[65] = .16$, *ns*). Likewise, a significant interaction between the two factors emerged ($F[2,130] = 18.28$, $p < .01$, $\eta^2 = .22$). Repeated contrasts indicated that group 1 scores differed between t1 and t2. Group 2 scores differed between all times of measurement ($p < .05$).

Difference scores were computed by subtracting t1 scores from t3 scores for all instruments. Differences scores of the self-assessment scale correlated significantly with difference scores of the knowledge test ($r = .23$, $p < .05$), and the difference scores of the search task procedure scores ($r = .36$, $p < .01$), while the correlation with the difference scores of the search task outcome scores did not reach significance ($r = -.02$, *ns*).

Hypothesis 3

For the examination of Hypothesis 3, correlations between the self-assessment scale, and the objective performance measures were computed for all times of measurement. At t1, the correlation with the multiple choice test was $r = .46$ ($p < .01$); the correlation with the search task outcome scores was $r = .11$ (*ns*), and the correlation with search task procedure scores was $r = .45$ ($p < .01$). At t2, the correlations were $r = .68$, $r = .33$, and $r = .55$ (all $p < .01$). At t3, all correlations dropped to nonsignificant values of $r = .09$, $r = .12$, and $r = -.01$.

Discussion

Hypothesis 1 is supported by the findings. Scores on objective performance measures increase as a consequence of participating in the instruction program. An important result is that group 1 outperforms group 2 at t2 on all measures (statistically, this is shown by the significant interactions between the time and the group factors). This shows that performance on the instruments does not simply increase incidentally or due to testing effects (repeated completion of the same task, or similar tasks, cfHausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007). In some cases, however, performance increased without participation in the instruction program (e.g. group 1 multiple choice test scores increased between t2 and t3, even though this group had participated in the instruction program earlier). One assumption is that the participants' interest in information literacy had been aroused by the instruction program or by completing the instruments, so participants possibly worked on information literacy materials on a voluntary basis.

Hypothesis 2 was also confirmed: Self-assessed information literacy increased by participating in the instruction program confirming the findings by Arnold et al. (1985) and Cassidy and Eachus (2000). Similar to the objective performance measures, group 1 scores were higher than group 2 scores at t2.

Further support for Hypothesis 2 comes from the significant correlations between increases on the self-assessment scale and increases on the knowledge test and search task procedure scores (in terms of difference scores). The correlation with increase on the search task outcome scores, however, was not significant. It can be assumed that the procedure scores are the more appropriate indicator of search task performance, as they indicate whether the participant was able to choose the most adequate information resource and used it adequately. Whether appropriate publications were found (which is measured by the outcome scores) depends not only on the correct procedure, but also on a variety of confounding factors. It is, for example, possible that a student chooses an unsophisticated procedure (e.g. just entering a few terms into Google Scholar) and is provided with one or two suitable publications right on the first results page. Another student might decide to use a bibliographic database (a far more sophisticated procedure for most search tasks), but fails to use a filter function. This student will most likely not find appropriate publications, though his performance level might be higher.

Hypothesis3 referred to the correlations of self-assessed and actual performance. Results show that self-assessed information literacy correlated significantly with the multiple choice test and the search task procedure scores at t1. In line with previous findings (e.g. Falchikov& Boud, 1989; Zell & Krizan, 2014), correlations were in the moderate range. However, the correlation with the search task outcome scores was not significant. As outlined above, it is likely that the procedure scores are the more appropriate indicator for performance on the search tasks.

At t2, correlations of the self-assessment scale and all objective performance measures reach significance level; yet, this can easily be explained, as t2 data are distorted due to differing training levels of the groups. Findings from t3 are interesting, as none of the correlations reaches significance. This is contrary to our hypothesis according to which self-assessment accuracy was expected to increase with performance level. Though of course, reduced variance might be an explanation for this finding, it seems unlikely to attribute this finding exclusively on this.

Additionally, it seems plausible to attribute this finding on the nature of the self-assessment scale: wording of the items was quite general to make sure that participants would be able to understand the items already before participating in the training.

After participation, however, participants had a more sophisticated understanding of the different aspects of information literacy. It can be assumed that the generally worded items of the self-assessment questionnaire were no more suitable for the participants, which means that the questionnaire lacked the required specificity for self-assessment (Mabe & West, 1982). This is in line with research on the related concept self-efficacy which has demonstrated that prediction of performance is better when using self-efficacy questionnaires specific to the domain or task (Bandura, 2006; Maddux, 1995). The assumption that the self-assessment scale is unsuitable for a population with a high performance level is further supported by the decreased reliability estimates at t3, as documented in Table 1. So, Hypothesis 3 could not be supported.

To summarize our findings, two out of three hypotheses could be confirmed. Apart from showing that the instruction program implemented lead to learning gains (Hypothesis 1), the results show that self-assessed ability increases along with objective performance measures (Hypothesis 2). It is especially striking to find that the self-assessment scores show the same change pattern as the objective performance measures in the sense that group 1 scores exceed group 2 scores at t2. Contrary to our expectations, Hypothesis 3 could not be supported. The correlations between the self-assessment scores and the objective performance measures failed to reach significance at t3 even though performance level was highest at this time of measurement. As explained above, this might be due to the unspecific wording of the self-assessment items.

In conclusion, an important implication for practice is that finding the right wording of a self-assessment scale can be difficult, especially when the scale has to be used with a population with great variability in performance level. It might be a solution to use more specific items than we did in our study and to investigate whether participants with a low performance level are able to understand and answer these items adequately.

Some potential problems or limitations should not be withheld. First, the findings described are based on a relatively small sample and limited to the domain of information literacy. Though, as most findings are in line with those from other studies that were conducted in different contexts, it can be argued whether the findings are generalizable.

Second, for the examination of Hypothesis 2, difference scores were used. This can be considered problematic for two reasons. (1) Calculating the difference between t3 and t1 scores for all participants ignores the group structure of the dataset. As the two groups participated in the instruction program in different periods, the difference between t3 and t1 scores might not be the same for both groups. (2) The use of difference scores is problematic per se due to the reduced reliability of these scores. The reliability of difference scores is strongly influenced by the correlation between the two components: the higher the correlation, the lower the reliability of the difference score (Cronbach & Furby, 1970; Peter, Churchill, & Brown, 1993). In the current study, though, correlations between the components of the difference scores were below $r = .40$; for the search tasks, the correlations between the components were even lower. Therefore, we do not consider low reliability of difference scores a severe problem in our study. Nonetheless, it would be more appropriate to use a different statistical approach to examine change patterns, like hierarchical linear modeling or latent growth modeling; however, the small sample size restricted analysis options.

Third, it should be addressed that Cronbach's Alpha and split-half reliability estimates of the multiple choice test were below the conventional criteria of .70 (Schmitt, 1996). Contrastingly, Guttman's Lambda 6 exceeded the .70 cutoff at all times of measurement.

Lambda 6 was computed in addition to the conventional estimates because information literacy is a heterogeneous domain (Peter, Leichner, Mayer, & Krampen, 2015) and Lambda 6 is a better reliability estimate if the test items are heterogeneous (Revelle, 2015, p. 230).Therefore, we think that the Lambda 6 estimate should be given more weight.

Two directions for future research that would complement the findings of this study should be mentioned. First, the data collected in this study do not allow examining overestimation or underestimation of performance. For examining this issue, it would be necessary to ask participants to estimate their performance in terms of the raw scores. As the instruments were new to the students and they did not receive any feedback on their performance, this was not possible to examine in our study. This might be an issue of further research. Another issue for future research might be examining whether self-assessed performance can explain variance in real-life performance beyond objective performance measures. This would require collecting performance data on a real-life task (i.e., references section of term papers).

To sum up, aside from showing the effectiveness of our instruction program, this study includes interesting findings related to the relationship between self-assessed and actual performance. Findings confirm that students are moderately able to self-assess their information literacy skills. Additionally, self-assessed abilities increase along with objective performance measures while participating in the instruction program. Contrary to the hypothesis and the findings from similar studies, the relationship between self-assessed and actual performance decreased with overall performance levels what might be attributed to the unspecific nature of the self-assessment questionnaire. So, our findings demonstrate the importance of adequate wording of self-assessment scales.

Acknowledgement: This work was supported by the German Joint Initiative for Research and Innovation with a grant acquired in the Leibniz Competition 2012 [grant number: SAW-2012-ZPID-6].

References

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*(4), 355–365. doi:10.1037/0003-066X.51.4.355

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

Arnold, L., Willoughby, T. L., & Calkins, E. V. (1985). Self-evaluation in undergraduate medical education: A longitudinal perspective. *Academic Medicine, 60*(1), 21–28.

Association of College and Research Libraries (ACRL). (1989). *Presidential committee on information literacy: Final report*. Retrieved from http://www.ala.org/acrl/publications/whitepapers/presidential

Association of College and Research Libraries (ACRL).(2010). *Psychology information literacy standards*. Retrieved from http://www.ala.org/acrl/standards/psych_info_lit

Assor, A., & Connell, J. (1992). The validity of students' self-reports as measures of performance affecting self-appraisals. In D. H. Schunk& J. L. Meece (Eds.), *Student perceptions in the classroom*(pp. 25-47). Hillsdale, NJ: Lawrence Erlbaum.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Adolescence and Education: Vol. 5. Self-Efficacy and Adolescence* (pp. 307–337). Greenwich, CT: Information Age Publishing.

Bandura, A. (2010). Self-efficacy. In I. B. Weiner & W. E. Craighead (Eds.), *TheCorsini Encyclopedia of Psychology*. New York, NY: John Wiley & Sons.

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education, 26*(1), 87–122.doi:10.1007/s12528-013-9077-3

Blatchford, P. (1997). Students' self assessment of academic attainment: Accuracy and stability from 7 to 16 years and influence of domain and social comparison group. *Educational Psychology, 17*(3), 345–359. doi:10.1080/0144341970170308

Bowles-Terry, M. (2013). Library instruction and academic success: A mixed-methods assessment of a library instruction program. *Evidence Based Library and Information Practice, 7*(1), 82–95.

Cassidy, S., &Eachus, P. (2000).Learning style, academic belief systems, self-report student proficiency and academic achievement in higher education. *Educational Psychology, 20*(3), 307–320.doi: 10.1080/713663740

Chamorro-Premuzic, T., & Furnham, A. (2006).Self-assessed intelligence and academic performance. *Educational Psychology, 26*(6), 769–779.doi: 10.1080/01443410500390921

Chen, H.-L. (2009). An analysis of undergraduate students' search behaviors in an information literacy class. *Journal of Web Librarianship, 3*(4), 333–347. doi:10.1080/19322900903328807

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68–80. doi:10.1037/h0029382

Dunn, K. (2002). Assessing information literacy skills in the California State University: A progress report.*The Journal of Academic Librarianship, 28*(1-2), 26–35. doi:10.1016/S0099-1333(01)00281-6

Eid, M., &Diener, E. (2006).The need for multimethod measurement in psychology. In M. Eid& E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 3–8). Washington DC: American Psychological Association.

Falchikov, N., &Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430. doi:10.3102/00346543059004395

Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education, 7*(2), 95–105. doi:10.1016/j.iheduc.2004.02.001

Grafstein, A. (2002). A discipline-based approach to information literacy.*The Journal of Academic Librarianship, 28*(4), 197–204. doi:10.1016/S0099-1333(02)00283-5

Grassian, E. S., & Kaplowitz, J. R. (2001).*Information literacy instruction: Theory and practice*. New York, NY: Neal-Schuman.

Haladyna, T. M., & Rodriguez, M. C. (2013).*Developing and validating test items*. New York, NY: Routledge.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability.*Journal of Applied Psychology, 92*(2), 373–385. doi:10.1037/0021-9010.92.2.373

Homann, B. (2003). German libraries at the starting line for the new task of teaching information literacy.*Library Review, 52*(7), 310–318. doi:10.1108/00242530310487407

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. doi:10.1037/0022-3514.77.6.1121

Kurbanoglu, S. S., Akkoyunlu, B., & Umay, A. (2006).Developing the information literacy self-efficacy scale.*Journal of Documentation, 62*(6), 730–743. doi:10.1108/00220410610714949

Leichner, N., Peter, J., Mayer, A.-K., & Krampen, G. (2013).Assessing information literacy among German psychology students. *Reference Services Review, 41*(4), 660–674. doi:10.1108/RSR-11-2012-0076

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

Levy, Y. (2007).Comparing dropouts and persistence in e-learning courses.*Computers & Education, 48*(2), 185–204. doi:10.1016/j.compedu.2004.12.004

Lew, M. D., Alwis, W., & Schmidt, H. G. (2010).Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education, 35*(2), 135–156. doi:10.1080/02602930802687737

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280–296. doi:10.1037/0021-9010.67.3.280

Maddux, J. E. (1995). Self-efficacy theory.An introduction. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment. Theory, research, and application* (pp. 3–33). New York, NY: Plenum Press.

McCulley, C. (2009). Mixing and matching: Assessing information literacy. *Communications in Information Literacy, 3*(2), 171–180. Retrieved from http://digitalcommons.linfield.edu/cgi/viewcontent.cgi?article=1001&context=librariesfac_pubs

Mery, Y., Newby, J., & Peng, K. (2012). Why one-shot information literacy sessions are not the future of instruction: A case for online credit courses. *College & Research Libraries, 73*(4), 366–377.

Moore, M. G., & Kearsley, G. (2012).*Distance education: A systems view of online learning* (3rd ed.). Belmont, CA: Wadsworth.

Nistor, N., & Neubauer, K. (2010). From participation to dropout: Quantitative participation patterns in online university courses. *Computers & Education, 55*(2), 663–672. doi:10.1016/j.compedu.2010.02.026

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

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), 969–983. doi:10.1002/asi.21030

Orr, R., Williams, M. R., & Pennington, K. (2009).Institutional efforts to support faculty in online teaching.*Innovative Higher Education, 34*(4), 257–268. doi:10.1007/s10755-009-9111-6

Peter, J. P., Churchill, G. A. Jr., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of Consumer Research, 19*(4), 655–662. doi:10.2307/2489447

Peter, J., Leichner, N., Mayer, A.-K., & Krampen, G. (2015).Making information literacy instruction more efficient by providing individual feedback. *Studies in Higher Education*. Advance online publication. doi:10.1080/03075079.2015.1079607

Pinto, M. (2010). Design of the IL-HUMASS survey on information literacy in higher education: A self-assessment approach. *Journal of Information Science, 36*(1), 86–103. doi:10.1177/0165551509351198

Revelle, W. (2015).*An introduction to psychometric theory with applications in R* [work in progress]. Retrieved from http://www.personality-project.org/r/book/

Romero, M., & Barberà, E. (2011).Quality of e-learners' time and learning performance beyond quantitative time-on-task.*The International Review of Research in Open and Distance Learning, 12*(5), 125–137. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/999

Rosman, T., Mayer, A.-K., & Krampen, G. (2014).Combining self-assessments and achievement tests in information literacy assessment: Empirical results and recommendations for practice. *Assessment & Evaluation in Higher Education, 40*(5), 740–754. doi:10.1080/02602938.2014.950554

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

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353. doi:10.1037/1040-3590.8.4.350

Skeff, K. M., Stratos, G. A., & Bergen, M. R. (1992). Evaluation of a medical faculty development program: A comparison of traditional pre/post and retrospective pre/post self-assessment ratings. *Evaluation & the Health Professions, 15*(3), 350–366. doi:10.1177/016327879201500307

Tung, R. (2010). *Including performance assessments in accountability systems: A review of scale up efforts*. Retrieved from http://files.eric.ed.gov/fulltext/ED509787.pdf

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

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science, 9*(2), 111–125. doi:10.1177/1745691613518075