Anette HIEMISCH[1] (Greifswald)

The Validity of Self-Assessments of Competences in Academic Course Evaluation

Abstract

The validity of self-assessments of competences may depend on the competence level of the respondents. Sixty-six psychology students evaluated the module methods with a test in which they assessed the competences they acquired in higher education. The scales of knowledge processing and personal competence showed significant correlations with the grade on the final oral exam. However, the expected correlation between the grade and specific competences (convergent validity) was only observed in a sample that excluded students with the lowest grades from the analysis. The results indicate that the self-assessments of less competent students may be more miscalibrated than those of more competent students; thus, the use of mean ratings may be problematic for the comparison of academic courses.

Keywords

academic course evaluation, self-assessment of competences, validity

Die Validität selbsteingeschätzter Kompetenzen im Rahmen der Lehrevaluation

Zusammenfassung

Hängt die Validität selbsteingeschätzter Kompetenzen vom Kompetenzniveau der Befragten ab? Sechsundsechzig Studierende der Psychologie evaluierten mittels eines Test zur Selbsteinschätzung erworbener Kompetenzen das Modul Methodenlehre. Die Skalen Fachkompetenz und Personalkompetenz korrelierten signifikant mit der Note in der mündlichen Abschlussprüfung. Das erwartete Korrelationsmuster zwischen Note und Kompetenzen (konvergente Validität) fand sich jedoch lediglich in einer Teilstichprobe ohne die Studierenden mit den schlechtesten Noten.
Die Selbsteinschätzungen kompetenter Studierender scheinen valider zu sein als die weniger kompetenter Studierender, wodurch die Nützlichkeit der Werte für den Vergleich von Lehrveranstaltungen oder Studiengängen eingeschränkt ist.

Schlüsselwörter

Lehrevaluation, Kompetenzerwerb, Selbsteinschätzungen, Validität

[1] hiemisch@uni-greifswald.de

1 Introduction

The focus on competences acquired throughout academic education implicates the need for a valid evaluation of learning outcomes and acquired competences (BRAUN & LEIDNER, 2010). This poses a series of methodological challenges different from those associated with the classic approach of student's evaluation of teaching. One of these problems is the validity of self-assessments of competences.

Over the last decade, an increasing body of empirical research has dealt with errors in self-assessment of competences or performance (e.g., BURSON, LARRICK & KLAYMAN, 2006; DUNNING, HEATH & SULS, 2004; KRUEGER & MUELLER, 2002; KRUGER & DUNNING, 1999). The findings cast doubts on the appropriateness of self-assessments for academic course evaluation. Consequently, the study addresses two questions: (a) are self-assessed competences valid indicators of the competences acquired throughout academic education, and (b) does the validity depend on the competence level of the students?

2 Self-Assessment of Competences

As the aim of academic course evaluation is to compare different courses or degree programs in regard to the competences acquired by the students, it is necessary to determine if errors in self-assessment of competences interfere with this aim. This is the case if: (a) self-assessments do not reflect the competence level of students, and (b) if self-assessments are flawed by systematic errors (e.g. students of a certain competence level systematically misjudge their competences), which implies that the mean ratings are an inappropriate measure for the comparison of courses.

2.1 Errors in Self-Assessment of Competences

In a seminal study, Kruger et al. (1999, p. 1121) found that students who were unskilled with regard to different competences (humor, grammar, logic reasoning) "grossly overestimated their relative test performance and ability". Apparently, this is caused by a lack of meta-cognitive skills. People with relatively poor abilities in a specific area do not possess sufficient knowledge to accurately assess the quality of their performance. As they fail to recognize their shortcomings, they conclude that they have done well.

The findings were replicated in a series of studies addressing competences with high ecological validity: Students were asked to predict their test performance (DUNNING, JOHNSON, EHRLINGER & KRUGER, 2003), or participants predicted their performance with a challenging exam, the preliminary rounds of a debate tournament, or a gun owner's knowledge of firearms at a Trap and Skeet competition (EHRLINGER, JOHNSON, BANNER, KRUEGER & DUNNING, 2008). In the latter study, participants made comparative as well as absolute assessments of their skills, and the retest reliability of the self-assessments was obtained to control for measurement errors. Moreover, incentives for accuracy were offered to motivate participants to provide accurate self-assessments.

Consistently participants with low skills highly overestimated their relative and absolute competence, while highly skilled participants underestimated their relative competences, albeit to a smaller degree. Incentives had no effect on accuracy. In accordance with the "unskilled – unaware" hypothesis, the errors of unskilled participants were based on misperceptions of the quality of their performance. In contrast, skilled participants had a fairly accurate impression of their performance, but misperceptions about their relative standing. They tended to think that others performed equally well.

2.2 Self-Assessment in Educational Contexts

Dunning et al. (2004) listed a series of educational methods that either impede or enhance accurate self-assessment. Teaching strategies that improve the acquisition of competences and enhance self-assessment are reviewing past performance and benchmarking. Both methods establish a quality standard, which is a necessary requirement for accurate self-assessment. Students learn through feedback what was right or wrong with a task and how to improve their performance; they also learn about their level of competence.

Hence, the same conditions that promote the lasting acquirement of competences also increase the validity of self-report assessment of these competences. Consequently, for self-assessments of academic competences to be meaningful, students must have (a) acquired the relevant competences to a sufficient degree, and (b) must have learned about their relative standing with regard to the competences.

2.3 Implications for the Evaluation of University Courses

Usually, a single course teaches specific knowledge, e.g. botanical systematics. However, for students to acquire related competences such as the classification of plants, additional courses have to build on the knowledge, and offer opportunities to practice relevant skills. Thus, a meaningful entity for the evaluation of competences should be a series of courses distributed over time and structured to teach certain knowledge and skills.

For the new BA/MA degree programs in Germany, this entity is a module. The description of a module specifies the learning objectives: Knowledge as well as competences. At best, the different courses comprising the module are coordinated to optimize the relevant learning opportunities. Although diploma degree programs are not organized in modules, similar entities can be identified. Consequently, modules should represent meaningful entities of evaluation, because, at best, they provide the preconditions for the acquirement of competences as well as for valid self-assessment.

Hence, in the following the results in a module-exam and in a test at the end of a single course are correlated with a series of self-assessed competences, based on either the competences acquired throughout the whole module or the single-course.

3 Method

Students of psychology assessed their competences in the subject methods (module) and in an introductory course to general psychology (single course).

Over three semesters, the module methods encompasses four lectures (two completed by written exams), three tutorials, one seminar, and one practical course. In the latter, students work together in small groups over one semester to plan, conduct, and analyze an experiment. Finally, they present the results in a written report and a poster or a talk. Students, thus, have the opportunity to acquire relevant knowledge, and practice the associated skills in the tutorials and the practical course. Furthermore, they receive different types of feedback on their performance. After the third semester, each student is tested in an oral exam that emphasizes integrative knowledge and lasts about 25 minutes.

The course introduction to cognitive psychology (in German "Einführung in die Allgemeine Psychologie I") teaches the basics of cognitive psychology (e.g., perception, attention, memory), scientific work (e.g., literature search, APA rules for citations), and learning as well as presentation techniques. The self-assessed competences acquired in the course were correlated with the results in a concluding multiple choice test.

3.1 Self-assessment of Acquired Competences

Self-assessments of competences were obtained with the HEsaCom, which "is designed to measure the self-reported acquisition of competences" (BRAUN & LEIDNER, 2010, p. 299). With 26 items, the acquisition of the following six competences was assessed: (1) Knowledge processing (six items, e.g., "As a result of this course, I can remember most of the important terms and facts from this course."); (2) systematic competence (three items, e.g., "This course has helped me organize my work."); (3) presentational competence (two items, e.g., "After presenting in this course I can structure my talks better."); (4) communication competence (five items, e.g., "This course has helped me speak more precisely."); (5) cooperation competence (five items, e.g., "My participation in the work group made it easier for me to stand up for constructive team spirit."); and (6) personal competence (five items, e.g.," I feel more inspired by the topics studied in this course than at the beginning.").

The reliability of the different scales is sufficient to good, as are the internal consistencies on the course level (BRAUN, 2008, Study 3). Significant correlations were found between HEsaCom scales and the ratings of the learning opportunities by trained observers (BRAUN, 2008, Study 5). It was demonstrated that the HEsaCom does not measure the same construct as classical measures of student's satisfaction (BRAUN & LEIDNER, 2010). Furthermore, a significant, but small correlation was obtained ($\gamma = -.11$) between the scale knowledge processing and the grades in a written exam (BRAUN, 2008, Study 7).

3.2 Participants

Two consecutive cohorts of psychology students were asked to evaluate the module methods. The 2010 cohort contained 56 students (aged between 20 and 37, M = 22.8, m = 6, f = 50), of these, 52 were registered for the exam and 41 participated in the study. In 2011, the cohort contained 59 students (aged between 19 and 42, M = 22.3, m = 5, f = 54), of which 53 were registered for the exam, and 25 participated in the study. Thus, 66 students provided complete data sets. In addition, 15 psychology students in the first semester who attended an introductory course provided evaluations of the course and the results in a concluding multiple-choice test.

All means were applied to grant anonymity to the participants. As the distribution of sex and age amongst psychology students allows the identification of students by these two indicators, participants were not asked to provide the information on the evaluation sheets. Furthermore, students were told that it was intended to correlate the evaluations with either the grade or the result in the test, so that they could make an informed decision if they wanted to provide the relevant data.

3.3 Evaluation

Both, the introductory course and the module were evaluated by the students with the HEsaCom.

3.3.1 Evaluation of the Module

At the end of the third semester, and one week (first cohort) or three weeks (second cohort) before the exam, students were asked to evaluate the module methods with the HEsaCom.

The questionnaires were distributed at the final session of the practical course. A smaller group, which attended the final session on another day, was contacted via e-mail and asked to download the questionnaire. The completed questionnaires could be posted into a sealed box in the department. In addition, questionnaires, a reminder of the study, and a sealed box were provided in the waiting area during the week of the exam, so that students could fill in the evaluation sheets while waiting for their exam.

In the questionnaire students were asked to provide a personal code. The term module was then explained and the courses constituting the module methods were listed. The term course was replaced with the term module for all HEsaCom items. The items were rated on a five-point scale, ranging from 0 (= not true) to 4 (= completely true).

The oral exam was conducted by the author who had no knowledge of whether students had filled in an evaluation form. After students had received their grade, they were reminded of the study. Each student received a sheet in which s(he) was to fill in his or her personal code and grade outside of the office and then to post it into the sealed box.

The possible grades, specified by the examination office, ranged from 1 (excellent) to 5 (not sufficient) with intermediate steps from 1.3 and 1.7 to 3.3 and 3.7.

3.3.2 Evaluation of the Introductory Course

Four weeks before the end of the seminar students evaluated the course. In the last session, students voluntarily participated in a multiple-choice test that consisted of two different parts: Fifteen multiple choice questions regarding knowledge and concepts of cognitive psychology, and three items regarding methods of scientific work.

3.4 Hypotheses

Other competences beside expertise (knowledge processing) are supposed to influence the performance in an oral exam: The better a student organizes and structures the workload and the topic (systematic competence), the better (s)he can prepare for the exam and determine the core issues of the subject. Communication competence should also enhance performance in an oral exam. Finally, personal competence is supposed to influence a student's engagement with the subject which, in turn, should promote self-directed learning processes and understanding. Accordingly, significant correlations (convergent validity) with the grade were expected for the respective scales. In contrast, no correlations (divergent validity) were expected for the presentational competence, as this is limited to presentations, and the cooperation competence, as this does not overlap with the requirements of an oral exam.

Although the "unskilled-unaware" hypothesis provides the theoretical background of the study, there are relevant methodological differences. It was not examined whether students participating in the study were able to correctly predict their grades or test results. In contrast, the grades represented the criterion for the validation of self-assessed competences. If, as expected, the validity of these self-assessments depends on the competence level of the students, higher correlations between the grades and the relevant competences were expected for competent than for less competent students.

In the case of the introductory course, knowledge processing, systematic competence and personal competence were supposed to influence the result of the test. However, as the conditions for meaningful self-assessment were less favorable than in the case of the module, substantially smaller correlations were expected.

4 Results and Discussion

Figure 1 depicts the overall distribution of grades compared to the distribution in the study. The most substantial differences between the two distributions were obtained for the grade 2.7 (only 5 of 15 students provided their grade), and for the grade 5 (only 1 of 5 students provided their grade).

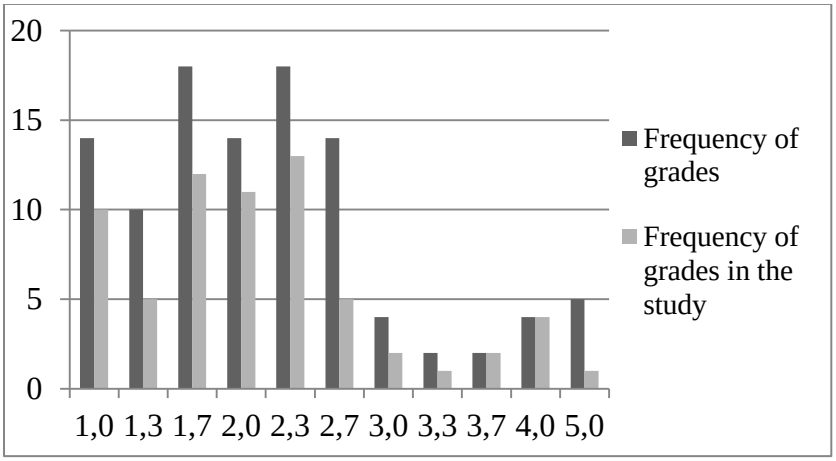

Fig. 1: Frequencies of grades

The reliabilities (Cronbachs's α) of the six HEsaCom scales ranged from satisfactory to excellent. Table 1 shows the descriptive statistics and reliabilities for all scales.

4.1 Evaluation of the Module

Spearman correlation coefficients were calculated between grades and competences for the whole sample, as well as for two subsamples: Sample12 excluded the 15 students whose grades were ≥ 2.7, while sample23 excluded the 27 students whose grades were ≤ 1.7. The results are depicted in Table 1.

In the complete sample, the scales knowledge processing and personal competence were significantly correlated with the grade, while only small and insignificant correlations were found for systematic and communication. As assumed, no correlations were obtained for presentation and cooperation competence.

As expected, in sample12 the pattern of correlations was even more in line with the expectations. Significant correlations were obtained for knowledge processing, systematic competence, and personal competence, while the correlation for communication competence was marginally significant. Thus, excluding students with lower competences from the analysis increased the validity of the ratings.

In contrast, in the sample that excluded the best students, the correlations between the grades and most of the self-assessments were practically zero. Exceptions were the scales systematic and personal competence. However, for systematic competences the prefix changed. That is, students with lower competences (higher grades) systematically overestimated their systematic competences.

Competences	*M*	*SD*	α	r(*N* = 66) $sample_{123}$	r(*N* = 51) $sample_{12}$	r(*N* = 39) $sample_{23}$
Knowledge processing	2.56	.68	.85	-.37**	-.41**	-.01
Systematic competence	2.35	.92	.83	-.10	-.26*	.20
Presentational competence	1.48	1.11	.90	.05	-.10	.05
Communication competence	1.74	.94	.90	-.10	-.23(*)	.01
Cooperation competence	3.07	.82	.89	.01	-.01	.02
Personal competence	2.37	.89	.87	-.30*	-.27*	-.25

* $p < .05$, ** $p < .01$, (*) p = .06, one-tailed test, $sample_{123}$ = complete sample, $sample_{12}$ = sample without the highest grades, $sample_{23}$ = sample without the lowest grades, the lower the grade the better the exam.

Table 1: Descriptive statistics, reliabilities and correlations between grade and scales

The results imply that, apart from personal competence, the ratings of students with lower skills were less valid than those of students with higher skills. The exception may be due to the fact that the reference point for personal competences is an internal state. Consequently, students might find it easier to provide the required self-report.

4.1.1 Evaluation of the Introductory Course

The correlations between the relevant HEsaCom scales and the points achieved in the two parts of the test (knowledge and scientific methods), as well as the descriptive statistics, are depicted in Table 2.

For knowledge processing negative correlations were obtained with both parts of the test, whereby the correlation with scientific methods became significant. However, as the criterion was the number of points, negative correlations imply that higher self-assessments of knowledge processing are associated with poorer test results. Systematic competence was also negatively correlated with the knowledge test. In contrast, for personal competence a negative correlation was obtained for the knowledge test, while a significant positive correlation emerged for the scientific methods test.

The evaluation of the single course was included into the study, to get an impression of the size of the correlations between the HEsaCom scales and a criterion under circumstances that did not promote the validity of the self-assessment. The results were in line with the expectations, as the self-assessed competences in regard of the course were less valid than those in regard of the module.

However, the large and systematic judgment error indicated by the negative signs was not expected. One possible explanation for the result is that students who judged their competence to be rather high did not find it necessary to prepare for the test and consequently achieved a low outcome. As this result is based on a small sample size, it has to be interpreted with care. On the other hand, in small seminars, the mean ratings may well be based on only 15 to 20 assessments.

	M	*SD*	Test results knowledge	Test results methods
Knowledge processing	1.79	.57	-.24	-.58*
Systematic competence	2.11	.73	-.29	.02
Personal competence	1.62	.54	-.20	.56*
Test result knowledge	25.33	6.59		
Test result methods	9.93	3.63		

* $p < .05$, one-tailed test

Table 2: Descriptive statistics and correlations between HEsaCOM scales and test results

5 General Discussion

Under circumstances that promote the validity of the self-assessment of competences, the validity of the scales knowledge processing and personal competence was supported. However, the correlation between knowledge processing and the grade increased when the students who did not perform well were excluded from the analysis, and disappeared when the students with the best grades were excluded. In addition, the expected pattern of correlations between the grade and specific HEsaCom scales was only observed in the sample that excluded the students who did not perform well. This pattern of results is in line with the "unskilled-unaware" hypothesis. Furthermore, on the course level high competence ratings were negative predictors of performance, i.e., students who rated their acquired knowledge as being high achieved a low outcome in the test.

In previous studies, correlations between self-assessment and objective performance lay between .29 and .35 (DUNNING et al., 2004). In comparison, the correlations obtained for the module were rather high. This is especially noteworthy as students were not asked to predict their grades, but to evaluate the acquisition of competences, so that the relation between the self-assessments and the criterion is only indirect.

Based on these results, it is possible to answer the questions raised above: Self-assessments provide valid measurements of competences, albeit only under specific circumstances. This constraint, however, renders the mean ratings problematic for the comparison of courses or modules, because it is impossible to tell, whether the mean ratings of a module are based on the self-assessments of rather competent students, or whether they are based on the unrealistic high ratings of students who are less competent. In the latter case, the ratings might be misleading and, therefore, are not useful as indicators of the quality of education.

5.1 Assessment of Competences

The implications for the self-assessment of competences are twofold. Firstly, to specify the requirements for a valid self-assessment of competences, a cognitive

model is necessary that describes the underlying psychological processes. To provide a correct self-report regarding a behavior, a respondent must at least (a) understand the question, (b) recall the relevant behavior, (c) correctly infer and estimate the frequency of the behavior, (d) map the answer onto the response format, and (e) edit the answer for reasons of social desirability (SCHWARZ & OYSERMAN, 2001). The cognitive processes underlying the self-report of competences are even more complicated. To assess his or her competences (e.g., regarding giving a talk), a student needs to possess relevant declarative (how to structure and present a talk) and procedural knowledge (experience of giving talks), as well as knowledge about different standards of performance and about his or her standing in relation to these (are my talks good and are they better or worse than those of other students). Based on this knowledge, (s)he must make a prediction about future performance that is not biased by optimism or social desirability. Due to the inherent complexity of this process, it is possible that, even under the best conditions, the validity of self-assessments might be problematic. Therefore, the most valid method to assess competences is still a direct evaluation of performance (as in an assessment center), an approach that is too costly for the evaluation of academic courses.

Secondly, to apply self-assessments of competences in academic course evaluation, it is necessary to check if the requirements for meaningful self-assessment are fulfilled. These seem to be the same factors that promote the lasting acquisition of competences. However, if the structure of a module, the provided learning opportunities, and the quality of feedback have to be assessed in order to determine the potential validity of the self-assessments, it may be more promising and economic to directly assess these factors as indicators of the quality of education. Thus, further research should investigate which aspects of courses/modules and of students' behavior are the best predictors of the acquired competences, and develop valid measurements of these predictors.

5.2 Limitations and Further Research

One objection to the "unskilled-unaware" interpretation of the results is that it depends on the adequacy of the grade as an indicator of competence. However, the specific pattern of results cannot be explained by generally calling the validity of grades into question. It must rather be assumed that the students in group three systematically received a grade that did not match their competence. While it is always possible to "misdiagnose" a student's competence level in an exam, such a systematic error seems unlikely. In addition, the difference in personal competence between group one and three indicates a systematic difference between the groups. Students who lack interest in a subject might engage less in the courses as well as in the preparation for the exam and thus perform worse.

Furthermore, it is possible that good grades are more valid than poor grades, because a variety of factors such as test anxiety, achievement motivation or bad health might affect the performance in an exam. Consequently, students might perform poorly despite existing competences. Therefore, further research should include these factors.

The results are based on relatively small samples. In this respect, they only represent a first step towards a comprehensive evaluation of the validity of self-assessed competences. Additional studies with larger samples and different validation criterions are necessary to fully understand the underlying psychological processes.

6 References

Braun, E. (2008). *Das Berliner Evaluationsinstrument für selbsteingeschätzte studentische Kompetenzen (BEvaKomp)* [Competence-based course evaluation: Berlin Evaluation Instrument for students' competences]. Göttingen: V&R unipress.

Braun, E. & Leidner, B. (2010). Academic course evaluation: Theoretical and empirical distinctions between self-rated gain in competences and satisfaction with teaching behavior. *European Psychologist, 14*, 297-306.

Burson, K. A., Larrick, R. P. & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology, 90*, 60-77.

Dunning, D., Heath, C. & Suls, J. M. (2004). Flawed Self-Assessment. *Psychological Science in the Public Interest (Wiley-Blackwell), 5,* 69-106.

Dunning, D., Johnson, K., Ehrlinger, J. & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*, 83-87.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D. & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*, 98-121.

Krueger, J. & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*, 180-188.

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*, 1121-1134.

Lucas, R. E. & Baird, B. M. (2006). Global Self-Assessment. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 29-42). Washington, DC: American Psychological Association.

Schwarz, N. & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation, 22*, 127-160.

Author

PD Dr. Anette HIEMISCH || Ernst-Moritz-Arndt Universität Greifswald, Institut für Psychologie || Franz-Mehring-Straße 47, D-17487 Greifswald

www.mnf.uni-greifswald.de/institute/institut-fuer-psychologie.html

hiemisch@uni-greifswald.de