Characteristics of effective exams—Development and validation of an instrument for evaluating written exams

Benjamin Froncek [a,*], Gerrit Hirschfeld [b,1], Meinald T. Thielsch [c,2]

[a] FernUniversität in Hagen, Department of Psychology, Universitätsstr. 33, 58097 Hagen, Germany
[b] Deutsches Kinderschmerzzentrum, Vestische Kinder- und Jugendklinik Datteln, Dr.-Friedrich-Steiner Str. 5, 45711 Datteln, Germany
[c] Westfälische Wilhelms-Universität Münster, Department of Psychology, Fliednerstr. 21, 48149 Münster, Germany

ARTICLE INFO

ABSTRACT

Students' feedback is common in teaching evaluation, but there is no documented instrument enabling instructors to systematically gather relevant student feedback on written exams in higher education. Three studies are described to develop a valid instrument for evaluating written exams. Study 1 analyzes characteristics of effective written exams from the perspective of students and instructors, using qualitative content-analysis. This informs study 2, which analyzes and revises the structure of a questionnaire via exploratory factor analysis. In study 3, confirmatory factor analysis and cross-validation are conducted to confirm the structure found in study 2. Central factors are "Transparency", "Composition of the exam" and "Students' workload". Students' feedback as assessed by this questionnaire provides reliable feedback to improve the quality of exams.

Introduction

Written exams, such as multiple-choice exams, are likely to be used as a summative assessment method, as they can assess many topics in one exam within a short amount of time, particularly in large classes (Fellenz, 2004). From an instructor's point of view, there are even more advantages to using multiple-choice-exams: "promptly available results; removal of marking error; and the banking of items for future use" (Ferrão, 2010, p. 821). Despite the ambivalent attitude of instructors and students toward multiple-choice exams, they are still viewed and used as an effective and efficient assessment method (Fellenz, 2004). Given the importance of the decisions that are based on the results of higher education exams (Ferrão, 2010), e.g. their profound effect on students' future careers, instructors should be highly competent in developing multiple-choice exams, but this is not always the case (Burton, 2005). Much guidance is given on how to prepare a multiple-choice question (see e.g., Burton, Sudweeks, Merril, & Wood, 1991; Clegg & Cashin, 1986). The criteria for effective written exams are often based on classic test theory, focusing on the overall difficulty or the internal consistency of the exam (Burton, 2005; Case & Swanson, 2001). Although most academic exams appear to be constructed adequately from this point of view (MacDonald & Paunonen, 2002), the question remains: Are these criteria sufficient for an effective exam, or do instructors need to take other aspects into account as well?

The Student Evaluation Standards (Joint Committee on Standards for Educational Evaluation, 2003) offer guidance on how to conduct student evaluations with propriety, utility, feasibility, and accuracy as constituent aspects of evaluations within educational settings. These important principles aid instructors in developing high-quality evaluation processes. Nonetheless, the Student Evaluation Standards are broad principles and it "needs to decide what it wants to be – a comprehensive vision of excellence for student assessment at all levels, or a document designed for classroom users of student assessment information" (Arter, 2009, p. 11). Thus, a valid evaluation instrument that is able to provide critical and specific information from the students' perspective is needed in higher education. Such an instrument should be based on a profound analysis of what constitutes a good written exam. Furthermore, this analysis should include the perspectives of students as well as instructors.

Thus, the aim of this paper is to describe the development and validation of such a questionnaire that gathers students' feedback on written exams, the Muenster Questionnaire for Evaluating Written Exams (in German: Muensteraner Fragebogen zur Evaluation von Klausuren, MFE-K). To the best of our knowledge, there are currently no instruments that explicitly ask higher

* Corresponding author. Tel.: +49 02331 987 2215.
  E-mail addresses: benjamin.froncek@fernuni-hagen.de
(B. Froncek), thielsch@uni-muenster.de (M.T. Thielsch).
  [1] Tel.: +49 02363 975 189.
  [2] Tel.: +49 02518 338 484.

education students for feedback on written exams. In the following discussion, we describe how we identified core aspects of effective written exams and how we developed and validated the MFE-K to assess these aspects.

What is an effective exam?

There are very few models describing criteria other than statistical criteria to describe and design effective written exams. Baartman, Prins, Kirschner, and Van der Vleuten (2007) proposed 12 quality criteria for a competence assessment program, which are summarized in the following section (adapted from Jonsson, Baartman, & Lennung, 2009).

Assessments should be accepted and agreed upon by all stakeholders (acceptability). Assessments should reflect competencies needed in future work situations (authenticity). Tasks should reflect required higher cognitive skills (cognitive complexity). The conditions within an assessment should be equal for all learners (comparability). The resources invested in assessment development and execution should be relative to its benefits (costs and efficiency). The consequences for learning and instruction must be considered (educational consequences). A proper mapping of exam requirements regarding the skills, knowledge, and attitudes at stake, excluding irrelevant variables, is important (fairness). Assessments should be in line with standards, curriculum, instruction, and assessment (fitness for purpose). Assessments should stimulate self-regulated learning (fitness for self-assessment). Assessments should be relevant for students and instructors (meaningfulness). Decisions made from the assessments should be accurate and constant across situations and assessors (reproducibility of decisions). Finally, students should realize and clearly understand the scoring criteria and the purpose of the assessment (transparency).

These criteria can be used for formative and summative assessments alike (Jonsson et al., 2009). Although they are related to competence assessments and were developed within a different setting, many of these aspects may also function as effective references for written exams. However, a problem with lists that use such a top-down approach to define effective exams is that it is unclear who can assess these characteristics; specifically, instructors cannot assess the degree to which specific aspects such as fairness are fulfilled.

Another line of research investigated assessments as part of students' approach to learning. That research employed the opposite approach by asking students about their perceptions of assessment. From that research we know that assessment is a defining feature of students' approaches to learning (see Entwistle, 1991; Ramsden, 1997). For example, students' concepts about the fairness of the assessment may influence their learning approaches and subsequent performance (Brown, 2011; Hirschfeld & Brown, 2009). Interestingly, these two approaches – one top-down, one bottom-up – agree on the importance of fairness, transparency and clarity.

The aim of the present research was to identify the core characteristics of effective written exams that are relevant to both instructors and students, and to develop a questionnaire that reliably assesses these aspects. To this end, we conducted a series of three studies. The first study used qualitative interviews with both instructors and students aimed at comprehensively assessing their concept of effective written exams. The second study used a first draft of the MFE-K questionnaire with a group of students and employed an exploratory factor analysis to test whether the proposed factors would emerge. The third study consisted of a cross-validation of the results of the second study, with a new sample using confirmatory factor analysis.

Study 1: qualitative analysis of the characteristics of effective exams

Method

Sample

Five students and five instructors participated in the first study. All were members of the Department of Psychology at the Westfaelische Wilhelms-Universitaet Muenster (WWU Muenster) in Germany. Students were in a bachelor program (B.Sc. Psychology), recruited from their first to third (and final) year of study. Four female students and one male student took part in the study. Instructors were long-term professors, all male and highly experienced teachers at the WWU Muenster.

Methods

A semi-structured interview format was used. The interview included two primary topics: (1) experiences of the interviewees with written exams, and (2) their concept of an effective written exam.

Participants were also asked to recount their best and worst experiences with written exams. The interviews lasted between 23 and 37 min, were recorded, and then transcribed. The transcribed interviews were analyzed by using qualitative content analysis (Mayring, 2000). This approach made it possible to identify categories of effective written exams from the collected data through inductive category development.

Results

Two independent and trained observers coded the data, one of whom was one of the authors of the present paper. Afterwards, intercoder reliability of the categories was calculated using Krippendorffs Alpha (Hayes & Krippendorff, 2007). In this analysis, Krippendorffs Alpha was .77 (95% CI: [.71, .83]), suggesting a fair intercoder reliability, since an Alpha of .80 would be sufficient and .67 would be required at the very least (Krippendorff, 2004). In addition, use of the bootstrapping procedure (Hayes & Krippendorff, 2007) ensured that the chance of the data being accepted as reliable when in fact they are not is quite low in our case (q = .007 for alpha$_{min}$ = .70).

The content analysis resulted in the creation of six categories (see Table 1). In total, 273 statements were extracted and taken into account. One notable result is that no outstanding differences between the two interviewee groups were found in the statements. Each group's description of the aspects was comparable in quality and approximately comparable in number.

The first category, "transparent requirements," describes a need for clarity regarding written exams. This category accounted for about 35% of all analyzed statements and is therefore the category most mentioned. Information about the procedural details of the exam (e.g., how much time there is, what type of support is allowed) should be given in advance. For this, preparatory activities, such as mock exams similar in type and length to the upcoming exam, were believed to be very helpful. It should also be explained how individual questions are weighted proportional to the entire exam.

The second category, "varying levels of difficulty", which covers about 18% of the statements, specifies that effective exams should vary in task complexity to differentiate student achievements. Tasks should also vary in terms of the type of knowledge required: Students should be able to show factual as well as applied knowledge. In this way, students can demonstrate their understanding and their ability to apply knowledge.

The third category, "layout", covering about 15% of the statements, describes the need for a good layout within exams.