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

Benjamin Froncek[1], Gerrit Hirschfeld[2], and Meinald T. Thielsch[3]
[1]Department of Psychology, FernUniversität in Hagen, Hagen, Germany
[2]University of Applied Sciences, Osnabrück, Germany
[3]Department of Psychology, University of Münster, Münster, Germany

Correspondence should be addressed to
Benjamin Froncek
Department of Psychology
FernUniversität in Hagen
Universitätsstr. 33
58084 Hagen
benjamin.froncek@fernuni-hagen.de

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 ex ams 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.

Key words: Teaching evaluation, Exam, Quality management, Higher education

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., Clegg & Cashin, 1986; Burton et al., 1991). 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 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 et al., 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 minutes, 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 $\text{alpha}_{\text{min}}$ = .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.

Table 1

Categorization of the characteristics of effective exams (based on the content analysis of 273 statements given in study 1).

Category	Sub-category	Illustrative statements
Transparent requirements	Given Information before the exam	"What is to be learned has to be clear and transparent." or "[…]something was tested, [and it was] totally unclear, that this would be tested."
	Preparation activities	"There are item examples, and they have to be valid."
	Weighting within exams	"[…] what score must be achieved to pass the test."
Varying levels of difficulty	Levels of difficulty	"One needs items that are both important and easy and items that are explicit and difficult."
	Knowledge	"One item should test different qualities of knowledge" or "testing something totally irrelevant is a shame."
	Application	"Knowledge, understanding and application, to get into some depth."
Layout	Examination grouping	"There is a clear structure that one can follow." or "A bad outline that doesn't tell, what's exactly being asked[…]."
	Examination buildup	"I like easier items at the beginning so I don't start to panic." or "Too many questions in too little time."
	Question formats	"I think it is good to have one part multiple-choice-items and another part open-ended questions."
Clarity	Question wording	"Items must be worded clearly; it is bad if one does not understand the question."
	Instructions	"I can ask questions during oral exams, but in written exams the instructions must be given within the test." or "Nested sentences in a question are inappropriate."
Consistency with the course content		"[…] an appropriate representation of the course content" or "subjects that are ambigous in learning, shouldn't be tested."
Ambient conditions		"Exams were accessible for all students at the same time."

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. Visual partitions, numbering, and themed sections support a fast and easy grasp of the exam tasks. To support students' workflow, easier tasks should be located at the beginning of the exam, or a mixture of multiple-choice and open-ended questions should be given.

The fourth category, "clarity", covering about 16% of the statements, describes the importance of well-worded questions and the intelligibility of the instructions to avoid misunderstandings. Instructions should be offered in an explicit way and without excessive descriptions.

The fifth category, "consistency with the course content", which covers about 9% of the statements, indicates that the content of the exam should provide a broad representation of all the content taught during the class, instead of testing a few selective topics at an inappropriate depth.

The sixth category, "ambient conditions", which covers about 7% of the statements, describes the need for similar and adequate room, light, and temperature conditions for each exam. All students should have the same amount of time to process the exam; seating arrangements or handing the exams out in envelopes can enhance the process.

Discussion study 1

The aim of the first study was to gather a description of effective exams from the perspective of both students and instructors. The analysis was carried out in a reproducible manner, all steps are documented and can be requested from the authors. It followed strict rules of coding, adhered closely to the raw data (Mayring, 2010), and reliability measures were conducted (Mayring, 2000), which means that quality criteria for content analysis are widely met. Since identifying the categories through inductive content analysis was done by one of the authors of this paper, this process cannot be presented as completely unbiased. However, having measured the intercoder reliability with an independent second coder, which yielded fair results, we believe that the categories are of substantial value for the content analyzed in this study.

We found that both groups had very concrete ideas about poor and good qualities of a written exam and were able to articulate these in some detail. Our content analysis of the interviews resulted in six categories into which these concepts could be divided. These categories partly overlap with existing ideas about effective exams. For instance, Burton

(2005) campaigned for exams that are clear and well-constructed and noted the importance of clear instructions. Several of the core aspects identified in the interviews were also similar to the criteria of a competence assessment program (Baartman et al., 2007), such as transparency, cognitive complexity, and comparability. In addition, there were several aspects that were not - or only implicitly - part of previous guidelines, such as layout, varying levels of difficulty, and clarity.

Although we only interviewed a small number of participants, we believe our findings reflect the entire range of relevant concepts because no novel aspects emerged in the final interviews. However, one limitation to the generalization of the findings of this study is that only members of the Department of Psychology at the WWU Muenster were interviewed. Further research is required to deepen the understanding of the core aspects of effective exams.

Study 2: Scale revision and exploratory analysis

Based on study 1, we widely revised an existing feedback form for written exams (a prior version of the MFE-K, see appendix) by creating several new items based on the categories found in study 1 and merging them with the former feedback form. We then investigated the individual item characteristics (difficulty, skewness, kurtosis) and used exploratory factor analysis to test the underlying factor structure.

Methods

Sample

The MFE-K was first used to evaluate eight written exams for the B.Sc. Psychology program in the 2010 summer term. In total, 525 questionnaires were returned (47.1 %.); of these questionnaires, 383 (73.0 %) were completed by females, 92 (17.5 %) were completed by males, and no gender was indicated on 50 (9.5 %) questionnaires. Age ranged between 18 and 55 years, and the mean age was 22.2 (SD = 3.4).

Materials

The new revised version of the MFE-K was used, including the items that were developed to assess the six categories that were found in study 1. The items covering the six categories offered a seven-point Likert scale as answer options, ranging from 1 (‘strongly disagree’) to 7 (‘strongly agree’).

The category “Transparent Requirements” encompassed four items: “The requirements were known before the exam”; “Content, subject areas and literature were known before the exam”; “The format of the exam (type of questions, length) was known before the exam”; and “Enough material for preparation (sample questions) was given before the exam”.

The category “Varying levels of difficulty” encompassed two items: “I think the

questions varied in difficulty", and "The exam was too difficult to me".
The category "Layout" encompassed two items: "I think the exam was well-arranged", and "I think the exam contained sufficient knowledge questions and practical questions".

The category "Clarity" encompassed three items: "The answer format caused me problems"; "I think the questions were precise and clearly worded"; and "I think the instructions were explicit".

The category "Consistency with the course content" encompassed two items: "I think the different parts of the course were reflected in the exam", and "What percentage of the content of the exam was taught in class?".

The category "Ambient conditions" encompassed three items: "The ambient conditions of the exam were appropriate (enough space, enough light, comfortable temperature, etc."; "It was quiet enough during the exam"; and "I was able to process the exam completely within the given time" (answer options: Yes | No).

The questionnaire consisted of newly developed items for the evaluation of the exam and the evaluation of students' preparation. The student workload was assessed with three items: "I had a hard time finding enough time to learn"; "Because of other exams I could not prepare myself properly"; and "The number of exams in this semester is a great burden to me". Three more items were included to control for the potential bias caused by students' previous knowledge, interest, or motivation (e.g., Olivares, 2001):

- "I have sought information about this exam by…" [answer options: examiner (during the course, office consultation, etc,) | other instructor | student tutor | students from my class | students from past terms | student council]
- "I am very interested in the subject." [answer options: seven-point Likert scale]
- "I only want to pass this test, I do not care about the grade." [answer options: Yes | No]

At the end of the questionnaire, students were asked to rate their satisfaction with respect to their performance ("I am satisfied with my performance in this exam" [answer options: seven-point Likert scale]). Two open-ended questions were added, one regarding problems in the test preparation phase and one for general comments for the instructors. Demographic information was assessed by three items (age, gender, and number of terms). The aim of study 2 was to analyze the factorial structure of this first draft of the MFE-K using explorative factor analysis.

Procedure

The questionnaire was given to the students directly after they had completed the exam. This was done in different courses and with different instructors. Starting on the day of the exam, students had two weeks to complete and return the questionnaire. Potential bias resulting from this procedure, such as lower average scores when experiencing difficult exams, are not to be expected, since students' experience with the exam showed no substantial effects on their ratings, as was shown by post-exam evaluations of teaching (Arnold, 2009).

Approximately 80% of the returned questionnaires were received on the day of the exam.

Data analysis

Negatively worded items were reverse coded before entered into the analysis. Data were analyzed using skewness and kurtosis, followed by an exploratory factor analysis (a principal axis analysis using oblimin rotation), which was performed on the remaining items.

Results

Item characteristics

The distribution of responses was skewed for most items. Four items particularly stood out in this analysis and were consequently transformed into dichotomous [Yes | No] scales: "I think the exam contained sufficient knowledge questions and practical questions"; "I think the different parts of the course were reflected in the exam"; "The ambient conditions of the exam were appropriate (enough space, light, comfortable temperature, etc.)"; and "It was quiet enough during the exam".

Answers to the other items of the questionnaire reflected values for skewness and kurtosis that are acceptable for factor analysis ($-1.58 \leq \text{skew} \leq 0.53$; $-1.19 \leq \text{kurtosis} \leq 3.04$). According to West, Finch & Curran (1995), factor analysis leads to interpretable results if the sample size was > 200 and items showed no skew > 2 or kurtosis > 7.

Exploratory factor analysis

After the analysis of item characteristics as described above, an exploratory factor analysis was calculated for the 13 remaining items. Excluded from this analysis were the two bias items: one item dealing with students' satisfaction with their performance and one item that asked for prior interest in the tested subject.

With a value of .86, the KMO-Test indicated the suitability of the data for factor analysis. Kaiser-criterion and parallel analysis suggested different numbers of dimensions (three vs. two). Since the three-factor solution yielded loadings that were easier to interpret, we choose this solution (see table 2). Five items constitute the first factor, which was named "Transparency" ($.44 \leq a \leq .95$). One item, "The format of the exam (type and number of questions) was known before the exam", showed small loadings on the second factor as well (.33). Due to theoretical considerations, we decided to keep this item in factor 1. Four items constitute the second factor, which could be named "Composition of the exam" ($.37 \leq a \leq .85$). Again, one item showed substantial loadings on another factor: "The answer format caused me problems" showed a small loading on the first factor (.31). It was also kept in factor 2 due to theoretical considerations. Three items constitute the third factor, "Students' workload" ($.54 \leq a \leq .88$). There were no cross-loadings for this factor. The item "I think the questions varied in difficulty" showed no loadings > .3 on any of the three factors and was therefore excluded from the questionnaire.

Examining the factor means, it seems that factor 1 and 2 have higher values than factor 3. Since the latter is the only factor where high values reflect student's difficulties, this actually seems to be a desirable result.

Table 2

Study 2: Items of the three scales including M, SD, Min, Max, $r_{it,}$ and Item-loading.

Item	Factor 1	Factor 2	Factor 3	M	SD	r_{it}
Cronbachs alpha	.82					
The requirements were known before the exam.	.95			5.27	1.65	0.74
Content, subject areas and literature were known before the exam.	.89			5.59	1.51	0.73
The format of the exam (type of questions, amount) was known before the exam.	.47	.33		5.63	1.60	0.61
Enough material for preparation (sample questions) was given before the exam.	.51			4.92	1.75	0.54
The exam was too difficult for me	.44			2.89	1.05	0.49
Cronbachs Alpha		.78				
The answer format caused me problems.	.31	.37		3.33	1.89	0.38
I think the questions were precise and clearly worded.		.85		4.92	1.59	0.69
I think instructions were explicit.		.73		5.23	1.57	0.68
I think the exam was well-arranged.		.68		5.33	1.52	0.63
Cronbachs Alpha			.77			
I had a hard time finding enough time to learn.			.79	3.66	1.83	0.63
Because of other exams I couldn't prepare myself properly.			.88	3.41	1.83	0.71
The number of exams in this semester is a great burden to me.			.54	4.64	1.83	0.49
Factors						
Factor 1	1	.68	-.28	5.17	1.38	-
Factor 2	.68	1	-.15	4.97	1.41	-
Factor 3	-.28	-.15	1	3.90	1.53	-

Note. Factor 1 = Transparency; Factor 2 = Composition of the exam; Factor 3 = Students' workload; Analysis is based on a total of n = 525 questionnaires from the summer term 2010; Factor-loadings less than .3 are not shown, the item "I think the questions varied in difficulty" showed no substantial loading and was excluded.

Discussion study 2

The aim of the second study was to investigate the characteristics of the MFE-K items and describe the internal structure of this measure. Analysis of the item characteristics showed that four items were better represented by a dichotomous answer format. For example, the item "The ambient conditions of the exam were appropriate (enough space, enough light, comfortable temperature, etc.)" was nearly always answered in a way one could only interpret as a clear "yes". In analyzing these data, we realized that such items are best given with a clear "yes/no"-option according to the recommendations of Lyons (1998). Thus, we removed those items from the scales, but not from the instrument – mostly due to administrative reasons. For example, it could be important for a faculty manager to see which room is not evaluated as appropriate for an exam, even if this concerns only one or two rooms. Thus, these four items were not included within the scales, but remained as dichotomous items within the instrument. Of the remaining 13 items, 12 items were grouped on three factors using exploratory factor analysis. Those factors were named "Transparency", "Composition of the exam", and "Students' workload".

Items that belonged to the categories "Consistency with the course content" and "Ambient conditions" assess directly observable behavior that could either be present or absent, so we decided to ask those questions in a dichotomous way in the future and provide feedback to the instructors. Aspects such as quietness were either present or not, and therefore could be asked and reported in such a way.

Overall, this structure is in line with the core aspects identified in study 1. The only difference is that items developed to tap into the categories "Clarity" and "Varying levels of difficulty" loaded onto the same factor, "Composition of the exam". The aspect of varying difficulty, which was mentioned in the first study, did not show a specific loading to a factor in the EFA. However, this aspect can be best assessed by direct measures of exam difficulty, such as the number of correct responses. In summary, study 2 found that several aspects may be assessed by using economical binary choice formats, and that three factors ("Transparency", "Composition of the exam" and "Students' workload") can be used to describe the responses across the remaining items. Doing so focuses students' general perceptions of written evaluations on three main scales.

Study 3: Cross-validation and confirmatory factor analysis

The aim of the third study was to cross-validate the proposed structure of the MFE-K in study 2 in a different sample using confirmatory factor analysis. We also wanted to quantify the discriminative validity of the measure. Discriminative validity refers to the ability of an assessment instrument to discriminate between groups or individuals (e.g., Haynes & O'Brien, 2000) and is considered an important aspect of criterion-related validity (e.g., Messick, 1980). Discriminative validity should not be confused with discriminant validity; the latter refers to the accuracy of an instrument in discrimination between targets. Concerning the MFE-K, discriminative validity refers to its suitability in distinguishing reliably between the evaluations of different written exams.

Methods

Sample

The analysis in study 3 was based on a total of n = 688 questionnaires from the 2010/2011 winter term. In the 17 written exams evaluated, the average return rate was 45 %. Of the questionnaires received, 104 (15.2 %) were returned by males, 547 (79.6 %) were returned by females, and 37 (5.2 %) were returned with missing gender information. Ages ranged between 18 and 48 years (M = 23.08; SD = 3.41). Most questionnaires (76.3 %) were completed by students enrolled in the B.Sc. Psychology program, 20.7 % by students in the M.Sc. Psychology program, 2.8 % by students in other programs, and 6 % were returned without information on the study program. Since all three of our studies were conducted in the same department, it is possible that some students took part in the evaluation of exams in both terms. Due to privacy reasons, we were not able to control for this issue. But there were no repetitions of exams of the same subject in the two assessed terms of study 2 and 3; we evaluated all the written exams from two different programs of study and five different age groups.

Materials

We used the same items to elicit the core dimensions as before. Thus, the final MFE-K questionnaire included 12 items grouped into three factors plus additional items (see table 3). In addition to the items used in study 2, two new items had to be added to this set for administrative purposes (Nos. 16 and 21).

Table 3
Final MFE-K with factors and corresponding items.

Corresponding factor	No.	Item[a]	Answer option
		age	open-ended
		gender	male | female
		number of terms	open-ended
Students' workload	1	I had a hard time finding enough time to learn.	7-point Likert scale[b]
Students' workload	2	Because of other exams I could not prepare myself properly.	7-point Likert scale
Students' workload	3	The number of exams in this semester is a great burden to me.	7-point Likert scale
	4	What problems did you have in preparation for this exam?	open-ended
	5	How many hours did you study for this exam?	open-ended
	6	I have sought information about this exam by…	examiner (during the course, office consultation, etc.) | other instructor | student tutor | students from my class | students from past terms | student council
Transparency	7	The requirements were known before the exam.	7-point Likert scale
Transparency	8	Content, subject areas and literature were known before the exam.	7-point Likert scale
Transparency	9	The format of the exam (type and number of questions) was known before the exam.	7-point Likert scale
Transparency	10	Enough material for preparation (sample questions) was given before the exam.	7-point Likert scale
Transparency	11	The exam was too difficult for me.	7-point Likert scale
Composition of exam	12	The answer format caused me problems.	7-point Likert scale
Composition of exam	13	I think the questions were precise and clearly worded.	7-point Likert scale
Composition of exam	14	I think instructions were explicit.	7-point Likert scale
Composition of exam	15	I think the exam was well-arranged.	7-point Likert scale
	16	I think the assessment load for this test was too high.	7-point Likert scale
	17	I am satisfied with my performance in this exam.	7-point Likert scale
	18	I am very interested in the subject.	7-point Likert scale
	19	I think the different parts of the course were reflected in the exam.	Yes | No
	20	I think the exam contained sufficient knowledge questions and practical questions.	Yes | No
	21	During the test, it was always clear to me how many points I could receive for each question.	Yes | No
	22	It was quiet enough during the exam.	Yes | No
	23	The ambient conditions of the exam were appropriate (enough space, enough light, comfortable temperature, etc.).	Yes | No
	24	I was able to process the exam completely within the given time.	Yes | No
	25	I only want to pass this test, I do not care about the grade.	Yes | No
	26	What percentage of the content of the exam was taught in class?	open-ended
	27	Explanatory notes for instructors (problems	open-ended

	regarding the exam, suggestions, commendations, criticism):

[a] Item-numbering in this table equates numberings within the MFE-K.
[b]Seven-point Likert scale, ranging from 1 (‘strongly disagree’) to 7 (‘strongly agree’).

Procedure

Data were collected in the same manner as in study 2.

Data analysis

Data were analyzed using confirmatory factor analysis. Discriminative validity was assessed with a multiple analysis of variance (MANOVA) with the exam as the between subject variable and the scales as dependent variables.

Results

Confirmatory factor analysis and scale analysis

Confirmatory factor analysis was used to describe the item covariance matrix and to determine whether results from study 2 could be replicated. Factors were allowed to correlate and each item was constrained to load on only one factor. A confirmatory model with three correlated factors led to an imperfect model fit ($\chi 2 = 267.27$, df = 51; TLI = .89, CFI = .92, RMSEA = .09, GAMMA = 0.94). Goodness-of-fit statistics dramatically improved after dropping one item from the “Transparency” factor (“The exam was too difficult for me”). Without this item the MFE-K demonstrated a good model fit ($\chi 2 = 107.83$, df = 41; TLI = .96, CFI = .97, RMSEA = .06, GAMMA = 0.97). Thus, with only one minor change we affirmed the expected three-dimensional structure.

Correlations between the three scales varied between -.30 and .62 and internal consistency of the three scales can be considered as fair to good, especially considering the shortness of all three scales (see table 4).

As in study 2, mean differences were apparent between the scales in that factor 3 had a lower score than factor 1 and 2.

Table 4

Study 3: Items of the final three scales including M, SD, Min, Max, $r_{it,}$ and Item-loading.

Item	Factor 1	Factor 2	Factor 3	M	SD	r_{it}
Cronbachs alpha	.84					
The requirements were known before the exam.	.89			4.82	1.66	0.76
Content, subject areas and literature were known before the exam.	.78			5.35	1.46	0.68
The format of the exam (type of questions, amount) was known before the exam.	.74			5.23	1.63	0.66
Enough material for preparation (sample questions) was given before the exam.	.63			4.30	1.90	0.59
Cronbachs Alpha		.76				
The answer format caused me problems.		.68		4.73	1.69	0.54
I think the questions were precise and clearly worded.		.79		4.66	1.59	0.65
I think instructions were explicit.		.77		5.30	1.39	0.61
I think the exam was well-arranged.		.56		5.65	1.27	0.47
Cronbachs Alpha			.79			
I had a hard time finding enough time to learn.			.86	4.14	1.83	0.70
Because of other exams I couldn't prepare myself properly.			.90	4.04	1.82	0.73
The number of exams in this semester is a great burden to me.			.49	4.34	1.68	0.48
Factors						
Factor 1	1	.62	-.30	4.92	1.39	-
Factor 2	.62	1	-.27	5.02	1.20	-
Factor 3	-.30	-.27	1	4.19	1.51	-

Note. Factor 1 = Transparency; Factor 2 = Composition of the exam; Factor 3 = Students' workload; Analysis is based on a total of n = 688 questionnaires from the winter term 2010/11; Factor-loadings less than .3 are not shown.

Discriminative validity

Discriminative validity is an important aspect of criterion-related validity. The critical question is whether the MFE-K can reliably distinguish between the different exams. To test this, we calculated a MANOVA with the 17 evaluated written exams as the independent variable and the three scales of the MFE-K as the dependent variables. The differences

between the evaluations of the different exams turned out to be highly significant with a strong overall effect size ($F = 11.88$; $df = 48$; $p < .01$; partial $\eta^2 = .22$). All three scales are sensitive to differences between the exams in question (see table 5). The largest differences were found for the "Transparency" scale (partial $\eta^2 = .385$).

Table 5
Discriminative validity of the MFE-K scales.

Dependent variable	df	F	Sig.	Partial Eta-square
Composition of the test	16	13.376	.000	.246
Transparency	16	25.725	.000	.385
Students' workload	16	8.921	.000	.178

Note. The results are from a MANOVA with written exams as the independent variable and the MFE-K scales as the dependent variable.

Discussion study 3

The aim of the third study was to confirm the factor structure of the questionnaire and to test the discriminative validity. We were able to confirm the factor structure suggested by study 2. Compared to their length, all three scales showed fair to high internal consistencies and may thus be used for a typical analysis on a group level.

Multivariate analysis with the exams as independent variables showed that the scales distinguish between different exams with large to moderate effect sizes. But one has to keep in mind that our results at this point only imply that the MFE-K clearly differentiates between exams. A proof of discriminant validity and the accuracy of this differentiation should be performed in future research, for example with a comparison with actual grades of students. Due to privacy issues we were not able to perform such an analysis as of now.

General discussion

In the present paper we reported the development of a new questionnaire to assess the characteristics of effective written examinations. Starting with qualitative interviews (study 1) to systematically capture the concepts of students and instructors alike, we were able to refine a paper and pencil questionnaire to make it easier to administer by converting several items into a checklist format (study 2), and were able to differentiate between different exams while maintaining a stable factor structure and fair internal consistencies (study 3). Thus, these results indicate that the MFE-K is a useful tool for providing feedback on exams. In the following section we discuss the three scales, "Transparency", "Composition of the exam", and "Students' workload," separately before turning to interpretations, limitations, and a general outlook.

The three scales

The first aspect of effective exams that we identified in this study describes the extent to which students are aware of the requirements before the exam. This awareness is expressed in ample information and sample exams. Furthermore, transparency can be achieved by providing information about how the various parts of the exam are interrelated and scored, as Baartman et al. (2007) advocated as part of their quality criteria for a competency assessment program. This is in line with the fact that learning is to a large degree influenced by assessments (Müller, 2011; Rindermann, 2009), and knowledge of these conditions is a necessary factor for students to successfully complete the exam (Jacobs, 1984). Thus, while it is not surprising that this awareness was a primary aspect, we are pleased that the scale that assesses this facet has good internal validity and can be used to discriminate between exams.

The second aspect we identified in the study describes the extent to which the layout follows a clear visual structure and supports the comprehension of the exam. This comprehension can be achieved by grouping questions, numbering, highlighting, etc. In line with this finding, Clegg and Cashin (1986) have provided several tips for a clear structure, such as listing options on separate lines. Furthermore, beginning each section with simple questions is a way of easing students into the exam. Duffield and Spencer (2002) argue that some of these aspects are related to fairness, which is in line with our finding of medium-sized correlations between the first and the second scale. When students find an exam unfair, they respond negatively to the exam (Peterson & Irving, 2008). Furthermore, this scale assesses how clearly individual questions and instructions are worded. For many instructors, clarity is a challenging aspect of developing an exam (Burton, 2005) and highlights the need for feedback. When we examined the exams evaluated in study 3 we found that there were large differences between them. Thus, we believe that students are very capable of providing a differentiated feedback with this MFE-K scale.

The third aspect we identified in this study describes the perceived burden of the individual exam. The scale was motivated primarily by administrative demands and the frequently voiced concerns regarding the increased workload for students as a result of the Bologna reform in Germany's higher education system (Dany, Szczyrba & Wildt, 2008). We found that this concern is not only an individual perception, but that the differences between different students' ratings of the same exam are larger than the differences between exams. Some students reported a high workload during their exam preparation, while others described no problems at all (Bechler & Thielsch, 2012).

Interpretation of the three scales

Regarding the interpretation of the three scales, we suggest describing results summarized in mean values. As all scales showed at least fair homogeneity, assessments based on the scales should be much more reliable than analyses based only on single items of the MFE-K. For the first two scales - "Transparency" and "Composition of the exam" - one should strive for high mean values, according to a simple rule like "the higher - the better". Transparent and well-organized exams are always very welcome, with positive effects not only on students' perception of exams, but probably also on their learning and subsequent performance (cp. Brown, 2011; Hirschfeld & Brown, 2009). The situation is different on the third scale,

"Student's workload", where low or medium values are desirable: Workload is the only factor where high values reflect difficulties. Additionally, this scale reflects fairly general aspects and the load of other exams as well, factors that are not fully controllable by only one specific instructor.

Moreover, a deeper understanding of the students' perception of exams within a department will result from the comparison of different exam evaluations within one subject. Especially analyses of the evaluation over time will be helpful not only for faculty administrators, but also for the instructors: by doing so, they can compare students' evaluations to changes and improvements made in the exams or the organization of the examinations.

Limitations and future research

While this research may be only a first step toward validating this scale, it is important to note that we share the same problems as other evaluations of teaching aspects: Namely, there are rarely external criteria against which these judgments may be validated (see Marsh, 1984). We attempted to avoid this problem by conducting qualitative interviews with the aim of capturing all aspects that are relevant and assessable by students. As a result, we believe that the content validity of the MFE-K is high. In addition, enhancing the MFE-K by asking instructors and students from the relevant department meets the need of it being "tied to its purpose" (Keeley, 2012, p. 14), which contributes to its content validity. Confirming the 3-factor model further supports the construct validity of the questionnaire. Both the internal consistency and the discriminatory power suggest that the MFE-K can be used as intended to reliably assess differences between written examinations.

However, some limitations must be kept in mind: The MFE-K has thus far been used only to evaluate written exams in psychology. Thus, an examination of a broader use of this instrument is needed. Furthermore, we suggest that MFE-K scale benchmarks with respect to different subjects be used to provide more feedback to the instructors and opportunities for comparison. Another limitation is the possibility of a bias caused by dependent data. This could happen if the same students or exams were repeatedly included in the data. We see this as a small source of potential bias because there were no repetitions of exams of the same subject in the two tested terms; we evaluated all the written exams from two different programs of study and five different age groups. Thus, only a few students repeatedly took part in our evaluations, but due to the anonymity of the evaluation we were not able to test for this bias.

Several follow-up studies could be derived from our results. First, using this tool to assess effective exams in other subjects may provide some insight into the differences between subjects. Another means of validating this measure would be to compare students' evaluations of written exams to their evaluations of teaching. Additionally, a comparison of MFE-K evaluations and actual grades of students would make it possible to investigate further aspects of criterion-related validity, such as discriminant validity and the accuracy of the MFE-K in distinguishing between different exams. Finally, we believe that the combination of ratings of a particular exam, the concepts of assessment (Brown, 2011), and actual achievement may be useful for developing a coherent model of student achievement.

Conclusion

We have reported the development and validation of a short questionnaire to gather students' feedback on written exams. Such an instrument was greatly needed: though exams are considered an important aspect of student learning and many instructors struggle with developing adequate exams that increase learning, no systematic way of receiving feedback on this subject was previously available. The developed questionnaire is based on aspects that are important to both instructors and students. Furthermore, the MFE-K contains those aspects of written exams that can be reliably assessed by students. Specifically, these are "Transparency", "Composition of the exam", and "Students' workload". Other aspects, such as the difficulty of single exam tasks, can be best assessed by direct measures of exam difficulty such as the number of correct responses. We hope that the MFE-K in its final form will provide useful feedback for both examiners and faculty administrators.

References

Arnold, I. J. M. (2009). Do examinations influence student evaluations? *International Journal of Educational Research, 48*, 215-224. doi: 10.1016/j.ijer.2009.10.001

Arter, J. (2009). Classroom assessment for student learning (CASL) perspective on the JCSEE Student Evaluation Standards. Paper presented at the annual meeting of the American Educational Research Association, San Diego in the Division H symposium JCSEE National Conference on Benchmarking Student Evaluation Practices. *http://ati.pearson.com/downloads/JCSEE%20Student%20Eval%20Practices%20Symp%20Paper%202009.pdf* (accessed november, 2013).

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & Van der Vleuten, C. P. M. (2007). Determining the quality of competence assessment programs: A self-evaluation procedure. *Studies in Educational Evaluation, 33*, 258–281. doi: 10.1016/j.stueduc.2007.07.004

Bechler, O. & Thielsch, M. T. (2012). Schwierigkeiten bei der Vorbereitung auf schriftliche Prüfungen [Difficulties in written exam preparation]. *Zeitschrift für Hochschulentwicklung, 7*, 137-156.

Brown, G. T. L. (2011). Self-regulation of assessment beliefs and attitudes: A review of the Students' Conceptions of Assessment inventory. *Educational Psychology, 31*(6), 731-748. doi:10.1080/01443410.2011.599836

Burton, J. B., Sudweeks, R. R., Merril, P. F., & Wood, B. (1991). How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty. *http://testing.byu.edu/info/handbooks/betteritems.pdf* (accessed january, 2012).

Burton, R. F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*, 65-72. doi: 10.1080/0260293042003243904

Clegg, V.L., & Cashin, W.E. (1986). Improving Multiple-Choice Tests. iDEA Paper No. 16. *http://www.theideacenter.org/sites/default/files/Idea_Paper_16.pdf* (accessed february, 2012).

Case, S., & Swanson, D. (2001). Constructing written test questions for the basic and clinical sciences. Philadelphia: National Board of Medical Examiners. *http://132.204.3.67/documents/pdf/mesure/reference/10.NBME_MCQ.pdf* (accessed february, 2012).

Dany, S., Szczyrba, B., & Wildt, J. (eds.) (2008). *Prüfungen auf die Agenda! Hochschuldidaktische Perspektiven auf Reformen im Prüfungswesen* [Exams on the agenda! Educational perspectives on reforms in auditing]. Bielefeld: W. Bertelsmann Verlag GmbH & Co. KG.

Duffield, K.E., & Spencer, J.A. (2002). A Survey of medical students' views about the purposes and fairness of assessment. *Medical Education, 36*, 879-886. doi: 10.1046/j.1365-2923.2002.01291.x

Entwistle, N. (1991). Approaches to learning and perceptions of the learning enviroment. Introduction to the special issue. *Higher Education, 22*, 201-204.

Fellenz, M. R. (2004). Using assessment to support higher level learning: the multiple choice item development assignment. *Assessment & Evaluation in Higher Education. 29*, 703-719. doi:10.1080/0260293042000227245

Ferrão, M. (2010). E-assessment within the Bologna paradigm: evidence from Portugal. *Assessment & Evaluation in Higher Education, 35*, 819-830. doi:10.1080/02602930903060990

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77-89.

Haynes, S. N., & O'Brien, W. H. (2000). *Principles and Practice of Behavioral Assessment.* Springer: New York.

Hirschfeld, G. H. F., & Brown, G. T. L. (2009). Students' Conceptions of Assessment: Factorial and Structural Invariance of the SCoA Across Sex, Age and Ethnicity. *European Journal of Psychological Assessment, 25*, 30-38. doi:10.1027/1015-5759.25.1.30

Jacobs, B. (1984). Ambiguitätsreduzierende Maßnahmen zum Abbau von Angst in der Prüfung [Ambiguity reducing arrangements for lowering fear of exams]. In K. Ingenkamp (Ed.), *Sozial-emotionales Verhalten in Lehr- und Lernsituationen.* Erziehungswissenschaftliche Hochschule Rheinland Pfalz.

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards: How to improve evaluations of students.* Thousand Oaks, CA: Corwin Press.

Jonsson, A., Baartman, L. K. J., & Lennung, S. A. (2009). Estimating the quality of performance assessments: The case of an 'interactive exam' for teacher competencies. *Learning Environ Res, 12*, 225-241. doi:10.1007/s10984-009-9061-z

Keeley, J. W. (2012). Choosing an Instrument for Student Evaluation of Instruction. In M. E. Kite (Ed.), Effective evaluation of teaching. A guide for faculty and administrators, 13-21. *http://teachpsych.org/ebooks/evals2012/index.php* (accessed may, 2012).

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411-433.

Lyons, W. (1998). Byond agreement and disagreement: the inappropriate use of Likert items in the applied research culture. *International Journal of Social Research Methodology, 1*(1), 75–83.

Marsh, H. W. (1984). Students evaluations of university teaching - dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754. doi:10.1037/0022-0663.76.5.707

Mayring, P. (2000). Qualitative content analysis. *Forum: Qualitative Social Research 1*, Art. 20. *http://www.qualitative-research.net/index.php/fqs/issue/view/28* (accessed september, 2011)

Mayring, P. (2010). *Qualitative Inhaltsanalyse: Grundlagen und Techniken [Qualitative content analysis: Principles and techniques]* (11. ed.). Weinheim: Beltz.

MacDonald, P., & Paunonen, S.V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*, 921-943. doi:10.1177/0013164402238082

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012–1027. doi:10.1037/0003-066X.35.11.1012

Müller, F. H. (2011). Prüfungen, ein eigenes Kapitel [Assessments, a chapter of its own]. *http://www.univie.ac.at/physik-didaktik/hochschuldidaktik/Vortrag090511.pdf* (accessed october, 2011).

Olivares, O. J. (2001). Student interest, grading leniency, and teacher ratings: A conceptual analysis. *Contemporary Educational Psychology, 26*, 382–399. doi:10.1006/ceps.2000.1070

Peterson, E. R., & Irving, S. E. (2008). Secondary school students' conceptions of assessment and feedback. *Learning and Instruction, 18*, 238-250. doi:10.1016/j.learninstruc.2007.05.001

Ramsden, P. (1997). The context of learning in academic departments. In F. Marton, D. Hounsell & N. J. Entwistle (Ed.), *The Experience of Learning: Implications for Teaching and Studying in Higher Education*, 198-217. Edinburgh: Scottish Academic Press.

Rindermann, H. (2009). *Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts* [Evaluation of teaching: introduction and overview of research and practice of course evaluation with an input of evaluating computer-based learning] (2. Ed.). Landau: Verlag Empirische Pädagogik.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modelling. concepts, issues, and applications*, 56–75. Thousand Oaks: Sage.

Appendix

MFE-K in an early version, summer term 2009.

No.	Item[a]	Answer option
	age	open-ended
	gender	male | female
	study course	B.Sc. Psychology | Psychology as minor subject
	number of terms	open-ended
1	I had a hard time finding enough time to learn.	7-point Likert scale[b]
2	Because of other exams I could not prepare myself properly.	7-point Likert scale
3	The number of exams in this semester is a great burden to me.	7-point Likert scale
4	I studied … weeks for this exam and prepared …hours per week on average.	open-ended
5	I have sought information about this exam by…	examiner (during the course, office consultation, etc.) | other instructor | student tutor | students from my class | students from past terms | student council
6	How many hours did you study for this exam?	open-ended
7	I am pleased with the organization of the exam.	7-point Likert scale
8	The requirements were known before the exam.	7-point Likert scale
9	I think the assessment load for this test was too high.	7-point Likert scale
10	I think instructions were understandable.	7-point Likert scale
11	The answer formats (open ended questions, multiple choice formats, etc.) caused me problems.	7-point Likert scale
12	The exam was too difficult for me.	7-point Likert scale
13	I am satisfied with my performance in this exam.	7-point Likert scale
14	I was able to process the exam completely within the given time.	Yes | No
15	The content of the test was taught during the lecture.	Yes | No
16	Explanatory notes for instructors (problems regarding the exam, suggestions, commendations, criticism):	open ended

[a] Item-numbering in this table equates numberings within the MFE-K.

[b]Seven-point Likert scale, ranging from 1 (‘strongly disagree’) to 7 (‘strongly agree’).