View metada

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

ePub^{WU} Institutional Repository

Margit Kastner and Barbara Stangl

Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter?

Article (Published)
(Refereed)

Original Citation:
Kastner, Margit and Stangl, Barbara (2011) Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter? *Procedia - Social and Behavioral Sciences*, 12. pp. 263-273. ISSN 18770428

This version is available at: http://epub.wu.ac.at/4691/
Available in ePub^{WU}: November 2015

ePub^{WU}, the institutional repository of the WU Vienna University of Economics and Business, is provided by the University Library and the IT-Services. The aim is to enable open access to the scholarly output of the WU.

This document is the publisher-created published version.

http://epub.wu.ac.at/

Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 12 (2011) 263–273

Procedia
Social and Behavioral Sciences

International Conference on Education and Educational Psychology (ICEEPSY 2010)

Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter?

Margit Kastner[a]*, Barbara Stangl[a]

[a]*Institute for Tourism and Leisure Studies,Vienna University of Economics and Business, A-1090 Vienna, Austria*

Abstract

Problem Statement: Nowadays, multiple choice (MC) tests are very common, and replace many constructed response (CR) tests. However, literature reveals that there is no consensus whether both test formats are equally suitable for measuring students' ability or knowledge. This might be due to the fact that neither the type of MC question nor the scoring rule used when comparing test formats are mentioned. Hence, educators do not have any guidelines which test format or scoring rule is appropriate.

Purpose of Study: The study focuses on the comparison of CR and MC tests. More precisely, short answer questions are contrasted to equivalent MC questions with multiple responses which are graded with three different scoring rules.

Research Methods: An experiment was conducted based on three instruments: A CR and a MC test using a similar stem to assure that the questions are of an equivalent level of difficulty. This procedure enables the comparison of the scores students gained in the two forms of examination. Additionally, a questionnaire was handed out for further insights into students' learning strategy, test preference, motivation, and demographics. In contrast to previous studies the present study applies the many-facet Rasch measurement approach for analyzing data which allows improving the reliability of an assessment and applying small datasets.

Findings: Results indicate that CR tests are equal to MC tests with multiple responses if *Number Correct* (*NC*) scoring is used. An explanation seems straight forward since the grader of the CR tests did not penalize wrong answers and rewarded partially correct answers. This means that s/he uses the same logic as *NC* scoring. All other scoring methods such as the *All-or-Nothing* or *University-Specific* rule neither reward partial knowledge nor penalize guessing. Therefore, these methods are found to be stricter than *NC* scoring or CR tests and cannot be used interchangeably.

Conclusions: CR tests can be replaced by MC tests with multiple responses if *NC* scoring is used, due to the fact that the multiple response format measures more complex thinking skills than conventional MC questions. Hence, educators can take advantage of low grading costs, consistent grading, no scoring biases, and greater coverage of the syllabus while students benefit from timely feedback.

© 2009 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and/or peer-review under responsibility of Dr. Zafer Bekirogullari of Y.B.

Keywords: Multiple Choice Test, Constructed Response Test, Many-facet Rasch Measurement;

1877-0428 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and/or peer-review under responsibility of Dr. Zafer Bekirogullari of Y.B. doi:10.1016/j.sbspro.2011.02.035

1. Introduction

Practically everyone gets evaluated today and students have to spend a lot of time and effort getting prepared for examinations and passing them. According to Crooks (1988) only sitting formal written tests takes nearly 15% of students' time. Furthermore, students have to cope with different examination types during their study. They might be examined orally, or they have to write a response to a series of questions or stated problems. Another possibility is that students have to demonstrate their knowledge in multiple choice (MC) tests. These different forms have in common that the questions have a stem which represents the question as a problem to be solved, but differ in either offering or not offering response alternatives.

For the so called constructed response (CR) questions examinees have to create their own answers which might be a short answer, an essay, a diagram, an explanation of a procedure, or the solution of a mathematical problem (Lukhele, Thissen, & Wainer, 1993). Even musical performances or visual arts portfolios are bundled under the term CR tasks (Pollack, Rock, & Jenkins, 1992).

An argument for the usage of CR questions in examinations most often mentioned is that it tests a deeper understanding of the subject material (e.g. Bacon, 2003; Rogers & Harley, 1999). Moreover, CR tests are the appropriate format to encourage students to critically analyze problems (Rotfeld, 1998), make applied decisions, and reflect changing social values (Katz, Bennett, & Berger, 2000). Additionally, CR items are more reliable, because compared to conventional MC tests guessing is minimized (Ruch & Stoddard, 1925) and examinees cannot derive the correct solution by a process of elimination (Gibbs, 1995). One disadvantage is that only relatively few questions can be included in tests which means that not all material taught is covered (Ventouras, Triantis, Tsiakas, & Stergiopoulos, 2010). Another problem mentioned by Zeidner (1987) is that students with poor writing skills are disadvantaged even if knowledge content is superior. Further drawbacks of CR exams are related to grading. Grading tend to be more subjective, despite established scoring criteria (Powell & Gillespie, 1990). Moreover, grading CR exams is time consuming (Ventouras et al., 2010) and a computerized evaluation of the answers is still problematic due to the fact that the "evaluation machine" does not understand the meaning. Therefore, it searches for keywords or matches the response given to a defined sample answer (Gibbs, 1995). Hence, these are reasons why CR exams are replaced by MC tests.

Contrary to CR questions MC questions have a stem and a list of possible answers from which test takers have to select (the) correct answer(s). There are different types of MC examinations. The most prominent MC test uses MC questions with one stem and some choices; one of these choices is correct, the other ones are incorrect alternatives, called distractors (Bradbard, Parker, & Stone, 2004; Jennings & Bush, 2006). Several other MC formats were designed to measure more complex thinking skills. The incorporation of more than one correct answer is a good way to reduce the chances of guessing a question correctly (Bush, 1999; Foster, 2005). These MC questions with multiple responses were already introduced by Dressel & Schmidt in 1953 (Ben-Simon, Budescu, & Nevo, 1997; Dressel & Schmid, 1953) but still, little research has been done on that question type (Ben-Simon et al., 1997; Berk, 1986).

Thanks to automated scoring, the grading costs of MC tests are low (Bennett, Ward, Rock, & LaHart, 1990). Exams are graded consistently and therefore, scoring biases do not exist, which means that cross–marking is not necessary (Farthing, Jones, & McPhee, 1998). This advantage is especially important for educational institutions under the pressure to handle large-scale examinations (Roediger & Marsh, 2005). Due to the greater fairness there is no room for debating about grades. Students benefit from timely feedback (Weiss, Gridling, Trödhandl, & Elmenreich, 2006) and from the possibility to sit exams at remote locations when the exams are conducted on the computer. Furthermore, in MC testing the writing speed of different students is not important (Farthing et al., 1998), and a larger amount of questions can be asked, which causes a greater coverage of the syllabus (Bennett et al., 1990; Walstad & Becker, 1994). On the other hand, tests presented in MC format pose multiple problems. MC questions are heavily influenced by the formulation of the questions and answers which might lead to cueing effects (Schulze & Drolshagen, 2006). Therefore, experienced developers and pre-tests are necessary to eliminate these problems. Additionally, MC questions only test isolated pieces of knowledge (Bennett et al., 1990). This might lead to the opinion that MC questions are not suitable for testing high-level thinking such as problem solving in a real-world context and are seen only as a proper instrument for trivial recognition of facts. In many cases question banks exist and MC questions are used more than once which might bias the results, in case students have access to old exams (Schulze & Drolshagen, 2006). Moreover, questions can be answered due to lucky guessing and even high marks are possible (Bush, 1999). Moreover, partial knowledge of students is ignored in many cases (Ben-Simon et al., 1997).

The study at hand aims to compare CR and MC questions with multiple responses since assessing an individuals' knowledge is of interest for educators and nowadays, MC tests are very common, and replace many CR tests. In the United States for instance many respectable tests use the MC format (American Marketing Association, 2001), and popularity is also stated for testing economics at the principles level (Chan & Kennedy, 2002). Nevertheless, there are some open problems this study addresses. First, are CR and MC questions with multiple responses tests equal? Further, which test format is easier for students? And, does it matter which scoring method is used for grading MC exams?

To answer these questions the paper first outlines the relevant literature regarding previous research on the comparison of CR and MC questions. Next, details concerning the methodology are given. The result section begins with the description of the sample, than similarities or differences of test formats are discovered and effects are revealed. Finally, the results are discussed and implications are drawn.

2. Theoretical Background

In literature there are several discussions regarding the (dis)similarity of different examination types such as MC and CR tests.

Some authors like Ackerman & Smith (1988), Van den Bergh (1990), or Wainer & Thissen (1993) conclude that CR and MC tests evaluate essentially the same and therefore, Walstad & Becker (1994) state that from an economical point of view

concerning time and cost reduction, MC tests should be preferred. Bacon (2003) adds that even if one likes to test high level knowledge, MC questions are an appropriate measurement tool.

In contrast, Becker & Johnston (1999), Anderson et al. (2000), as well as Dufresene, Leonard, & Gerace (2002) concur that a different dimension of knowledge is measured with MC and CR tests. According to Martinez (1999) and Hancock (1994) this view is only partly correct. For the higher levels of Bloom's taxonomy of learning which can be thought of as degrees of difficulties and a useful structure to categorize test questions (for further information see Bloom, 1956) they found that there is no overall equality. Only within the first four dimensions which are remembering, understanding, applying and analyzing, they stated that MC and CR measure the same level of knowledge. However, there is little knowledge on the different level of difficulty of MC items and CR items (Katz et al., 2000).

This inconstancy of result is also uncovered by a meta-analysis of Rodriguez (2003) who examined 67 empirical studies. A closer look on the studies reveals that stem-equivalent MC and CR tests are higher correlated as more dissimilar ones. Shepard (2008) confirms the similarity in case of stem-equivalent questions measuring mathematical competency. Additionally, the evaluated domain is important. Measures are for instance equivalent in mathematical reasoning tests no matter what test format is used (Traub & Fisher, 1977). This is not the case for tests of verbal comprehension, and CR tests are different to the other test formats in the study (Traub & Fisher, 1977). In'nami & Koizumi (2009) also shed light on the test effects of first and second language reading and listening performance in their meta-analysis, and conclude that MC tests are easier than CR tests in first language reading and second language listening while no effects are found for second language reading.

While literature on test format effects is enourmous, hardly anyone addresses the scoring method used for grading the MC test. One exception is the study by Traub & Fisher (1977) who compare CR tests with conventional MC tests where the correct answer has to be identified and Coombs' tests where all incorrect distractors have to be identified by examinees. The scoring of Coombs' tests is known as elimination scoring which allows distinguishing between full knowledge (elimination of all distractors), partial knowledge (elimination of some distractors), partial misinformation (elimination of the correct answer and some distractors), and full misinformation (elimination of the correct answer alone). Absence of knowledge means that either no alternative is marked or all of them are marked (Ben-Simon et al., 1997). Findings of Traub & Fischer's study (1977) are that CR tests differ from MC tests and Coombs' tests, although MC und Coombs formats generate equivalent measures. Other researchers such as Ward (1982) and Walstad & Becker (1994) at least reported the type of MC questions and the scoring method applied. Both used MC questions with only one correct answer and four distractors; for grading *Number Correct* (*NC*) scoring with a correction for guessing is used. This means that the number of correct answers are summed up and deducted by 1/(n-1) for each incorrect answer (n = answer-possibilities). Contrary to Traub & Fisher they did not detect a difference between CR and MC tests.

Literature reveals that there is no consensus whether both test formats are equally suitable for measuring students' ability or knowledge. This might be due to the fact that neither the type of MC question nor the scoring rule used when comparing test formats are mentioned. Hence, educators do not have any guidelines which test format or scoring rule is appropriate. These shortcomings are addressed in the following experiment and are the guideline for the study.

3. Method

A total of 13 graduate students from the Vienna University of Economics and Business participated in the experiment. For the experiment three instruments are used. First a CR, and second a MC test which are given on the same day in paper-and-pencil format. Prior the test students did not have access to example questions of neither test format. Concerning testing time there were no time constraints. Hence, writing speed do not limit the value of the study and the same amount of questions can be asked in CR and MC tests to cover the whole syllabus. This procedure enables the comparison of the scores students gained in the two forms of examinations. Third, a questionnaire was handed out one week later for additional insights.

The first instrument is a CR test including 17 questions on different complexity levels.[1] To ensure validity, fairness, and reliability of the CR test, the following recommendations for grading given by Hogan & Murphy (2007) were respected. Tests were scored without knowing the identity of the examinees. Further, the proposed solutions (sample answers) were developed and a key was designed to reward partially correct answers. Incorrect answers were not penalized. This key was used as a guideline during grading. The grading procedure was as follows: The examiner scored the first question of each examinee, than the examiner moved to the next item to score it. Moreover, an experienced lecturer graded the exam because they grade tests more consistent than inexperienced ones (Weigle, 1994).

The second measure is a MC version of the first test using a similar stem to assure that the questions are of an equivalent level of difficulty. The stem was adapted to include instructions how to complete the task. In order to measure more complex thinking skills and to minimize the chance of guessing, multiple response questions are used. To assure the quality of the MC test all MC questions were a) constructed by a very experienced instructor whose job is developing online material including MC questions in Marketing for the e-learning-system of the university, and b) pretested on another student cohort. Prior to the test, examinees were instructed to answer each question, and they were informed that each question has one or more correct answers. For reliable and objective results MC tests were scored automatically, using three different scoring rules. The first one is called *All-or-Nothing* (*AN*) *scoring rule* and to get the credit students need to find all correct matches. Otherwise they will receive a score of

[1] Some examples are displayed in the Appendix.

zero. This scoring has a disadvantage for students with low ability (Reid, 1976), because partial knowledge of students is not captured and guessing is not possible (Ben-Simon et al., 1997; Bereby-Meyer, Meyer, & Flascher, 2002). The second scoring rule is very common and rewards partial knowledge in case of MC questions with multiple responses. The so called *Number of Rights* or *Number Correct* (*NC*) *scoring rule* is very simple, because only the correct responses are counted while incorrect tags are ignored (Ben-Simon, et al., 1997). The next scoring rule in use is the *University-specific scoring rule*, abbreviated *WU*. This rule not only rewards partial knowledge, at the same time it prevents guessing by penalizing incorrect tags as follows: Each task has a maximum number of points (*max*), and there are some correct (*r*) and some false (*f*) answer alternatives (at least one alternative has to be correct). For each correct alternative identified r/max_r points will be assigned and for each false alternative marked f/max_f will be subtracted; negative scores are prevented due to the constraint task score ≥ 0 (Learn@WU, 2007).

Additionally, demographic aspects as well as test preferences, learning strategies, and test anxiety of the examinees are collected. For measuring learning strategy the two-factor Study Process Questionnaire by Biggs et al. (2001) is used which allows evaluating whether students use a surface or deep learning approach. The measurement of test anxiety is adapted from Driscoll (2004) while the extrinsic and intrinsic motivation measure is borrowed from Pintrich et al. (1991).

4. Analysis

Many researchers already investigated (dis)similarities of test formats using correlations. In contrast, the present study applies the many-facet Rasch measurement (MFRM) approach. MFRM enriches the basic Rasch model which was proposed by the Danish mathematician Georg Rasch in the 1960's. The basic Rasch model provides a sample-free measurement (Bühner, 2006) which can calibrate any person's ability and item's difficulty independently of each other (Bond & Fox, 2007). MFRM is designed to integrate additional "measurement facets" (Linacre, 1994, 2009b) that influence test scores, such as rater severity, or person's characteristics. As a further advantage, MFRM also analyses data collected by ranking or rating scales, and is not restricted to dichotomous data only (Linacre, 2009b).

The software FACETS developed by Linacre (2009a) is used to apply MFRM and separate FACETS analyses are run to compare the CR with the MC test using the three different scoring rules (*AN, NC, and WU scoring rule)* described in the Methodology section.

To determine whether one of the scored MC tests is of similar severity than the CR test, measurement properties of FACETS such as *Separation*, S*eparation Reliability*, and *Strata* are inspected. The *Separation (index)* is the distance of *logits* between tests of varying severity. The higher the value of the *Separation (index)*, which has a range from zero to infinity, the more spread out is the measurement facet along the measurement scale, and the better is the discrimination of the measurement facet (Fisher Jr., 1992). The S*eparation Reliability* is the Rasch equivalent to Cronbach's alpha and refers to the ability to differentiate between tests or other measurement facets (Linacre, 2009b). It ranges from 0 to 1 and Fox & Jones (1998) considered S*eparation Reliability* equal or greater than .80 as acceptable. The number of *Strata* is calculated using the formula $(4 \times Separation)/3$ (Linacre, 2009b). Exams are interchangeable and form a homogeneous group regarding their severity tendencies, if *Strata* is just more than 1. The separation reliability underpins the circumstance if it goes towards 0 (Eckes, 2003). To provide evidence of the degree to which the scores of the exam are internally self-consistent two item-fit statistics (infit and outfit) are calculated. The infit statistic is an information weighted mean-square residual which is sensitive to inlying deviation, while the outfit statistic is an unweighted mean-square residual sensitive to outliers. The expected value is 1, and ranges from 0 to infinity. According to Linacre (2002) acceptable fit values are between .5 and 1.5. Other researchers recommend as a rule of thumb a narrower range from .7 to 1.3 (Bond & Fox, 2007), although acceptable values vary across disciplines (Wright & Linacre, 2002). The overall data-model fit is assessed by examining unexpected responses. According to Linacre (1994, 2009b) a model is satisfactory if about 5% or less of (absolute) standardized residuals are outside ±2, and about 1% outside ±3.

5. Results

Sample Description

The sample consists of 13 graduate students attending a Marketing course. There are more female (62%) than male students (38%). The majority of the students (53%) have a full-time job and additional 29% work in part-time jobs. Hence, the youngest students are 24 years old and five students are between 30 and 47 years. Based on their former education, already 40% of the participants have a university degree.

The learning strategies students used to prepare for the test varied. Some students used surface learning as their learning strategy. This means that they only try to memorize things in order to be able to reproduce them during the exam. From a pedagogical point of view these students "miss the point" of learning which should be an understanding and relating of new information to previous knowledge and personal experience. This is also known as a deep learning strategy and 62% of the students employed this strategy. Although students have different preferences concerning test format (preference for CR: 38%; preference for MC: 31%; no preference: 31%), no correlation between the preference and the learning strategy can be detected. Moreover, the preference for a specific test format does not have any impact on the exam results. Students are rather intrinsically motivated and prefer new, challenging, exiting tasks. However, future career perspectives motivate them, too. Furthermore, test anxiety is not observed.

Comparison of CR and MC test

Results of the FACETS analysis comparing CR with MC tests using the *AN scoring rule* are visualized in Figure 1. Looking at Figure 1 from left to right the first column represents the Rasch measure (*logit*) and 0 indicates an average test, an average student, and an average question. The second column shows the severity/leniency of the test format. In the present analysis the CR test is easier for examinees. This means that students got higher scores for the CR test than for the MC test (*logit* for MC-*AN* .25 and for CR -.15). Thus, the two test formats are far from being alike. This is consistently revealed by the separation statistics: The χ^2 = 106.8 (*df* = 1) is highly significant ($p < .001$). Since the *Separation index* is higher than 1 tests are not similar in severity/leniency (*Strata* = 14.5, *Separation* = 10.29), and *Separation Reliability* attests a very high divergence (*R* = .99).[2] The next column shows the ability of students, whereas students with high ability are displayed at the top of the column (*logit* range from .32 to -.25). The column "questions" arranges the questions by their difficulty. The most difficult question is called "production" (*logit* = .27) and the easiest question is "ratios" (*logit* -.24). Note that some example questions are displayed in the Appendix.

Furthermore, the fit statistics for exams indicate the degree to which the scores of the exam are internally self-consistent. No matter what range (.5 to 1.5 or .7 to 1.3) is used, both tests are scored consistently with values between .84 and 1.22.

The overall data-model fit is assessed by examining unexpected responses. Observations just exceed reference values because 5.5% were associated with (absolute) standardized residuals outside ± 2, and 1.9% are outside ± 3.

```
+-------------------------------------------------------------------------------------------------+
|Measr|    test    | students | questions                                                        |
|-----+------------+----------+------------------------------------------------------------------+
|   1 +   severe   +   able   + difficult                                                         +
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            | *        | production                                                       |
|     | MC-AN      |          | budget1          budget2        research                         |
|     |            | ****     | forecasting      r&d                                              |
*   0 *            * ****     * diversification  portfolio                                        *
|     |            | *        | perceptual map   positioning    projects  segments  semantic scales |
|     | CR         | ***      | distribution     products       ratios                           |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|     |            |          |                                                                  |
|  -1 +   lenient  +  unable  + easy                                                              +
|-----+------------+----------+------------------------------------------------------------------+
|Measr|    test    | * = 1    | questions                                                        |
+-------------------------------------------------------------------------------------------------+
```

Fig. 1: FACETS results comparing CR with MC tests using *All-or-Nothing scoring rule*

Next, the CR test is compared with the MC test using the *WU scoring rule*. The graphical description of the variables is displayed in Figure 2. Results indicate that both tests are different, although the logit spread is smaller than before (*logit* for MC-WU -.04 and for CR -.22 compared to the *logit* for MC-*AN* .25 and for CR -.15). *Strata* (5.91) and *Separation* (4.18) also confirm that there is a difference between the two test formats, which is highly significant ($p < .001$) with a χ^2 of 18.5 (*df* = 1) and a high reliability of the separation index (*R* = .95). The *WU scoring* discriminates better between students' ability (*logit* range from .45 to -.29) than the *AN scoring rule*. The difficulty of questions paints a similar picture as in the first comparison, and the questions are able to discriminate between students with different knowledge or ability. Concerning infit and outfit statistics values are acceptable with values between .92 and 1.14.

Model fit is satisfactory because 3.6% were associated with (absolute) standardized residuals outside ± 2, and .7% are outside ± 3.

[2] Table 1 provides an overview of statistics of all separate FACETS analysis.

```
+-----------------------------------------------------------------------------------------------------+
|Measr|    test    | students | questions                                                            |
|-----+------------+----------+----------------------------------------------------------------------+
|   1 +   severe   +   able   + difficult                                                             +
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            | *        |                                                                      |
|     |            |          | production                                                           |
|     |            |          |                                                                      |
|     |            | *        | budget1          research                                            |
|     |            | ***      | budget2          forecasting                                         |
*   0 * MC-WU      * **       * diversification  portfolio       projects  r&d     semantic scales  *
|     |            | ***      | positioning      segments                                            |
|     | CR         | **       | distribution     perceptual map  products  ratios                     |
|     |            | *        |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|  -1 +   lenient  +  unable  + easy                                                                  +
|-----+------------+----------+----------------------------------------------------------------------+
|Measr|    test    | * = 1    | questions                                                            |
+-----------------------------------------------------------------------------------------------------+
```

Fig. 2: FACETS results comparing CR with MC tests using University-specific scoring rule

Figure 3 compares the CR test with the MC test which applied *NC scoring* for grading. Again in the second column the severity/leniency of the test format can be spotted and at a first glance they have the same severity. In detail *logit* measures are .22 for the MC test and .18 for the CR test.

```
+-----------------------------------------------------------------------------------------------------+
|Measr|    test    | students | questions                                                            |
|-----+------------+----------+----------------------------------------------------------------------+
|   1 +   severe   +   able   + difficult                                                             +
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            | *        |                                                                      |
|     |            |          |                                                                      |
|     |            |          | production                                                           |
|     |            |          | budget1          budget2       research                              |
|     |            | ****     | diversification  forecasting   r&d                                   |
*   0 *            | * **     * portfolio        projects      semantic scales                       *
|     |            | ***      | segments                                                             |
|     | CR   MC-NC | ***      | perceptual map   positioning   products         ratios               |
|     |            |          | distribution                                                         |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|     |            |          |                                                                      |
|  -1 +   lenient  +  unable  + easy                                                                  +
|-----+------------+----------+----------------------------------------------------------------------+
|Measr|    test    | * = 1    | questions                                                            |
+-----------------------------------------------------------------------------------------------------+
```

Fig. 3: FACETS results comparing CR with MC tests using *Number Correct scoring rule*

To address the interchangeability of test formats, one needs to examine severity measures. This information is given by the *Strata* of 1.17 and a *Separation Reliability* of .28 for the comparison of CR with MC tests using *WU scoring*. The reference values provided by Eckes (2003) show that these test formats are interchangeable and form a homogeneous group regarding severity because *Strata* is just more than 1 and *Separation Reliability* goes towards 0. Furthermore, $\chi^2 = 1.4$ ($df = 1$) is not statistically significant ($p = .24$), indicating that tests are similar which supports the observation in Figure 3.

Model fit is satisfactory because 3.4% were associated with (absolute) standardized residuals outside ± 2, and .7% outside ± 3.

Table 1: Summary Statistics for the Many-Facet Rasch Analysis

Statistics	Comparison CR & MC-AN	Comparison CR & MC-WU	Comparison CR & MC-NC
RMSE	.03	.03	.03
S.D.	.28	.12	.02
χ^2	106.8*	18.5*	1.4
df	1	1	1
Strata	14.05	5.91	1.17
Separation index	10.29	4.18	.63
Separation reliability	.99	.95	.28

* p < .001

6. Conclusion and Implications

In previous studies researchers hardly mention which MC questions or scoring rules are used when comparing test formats. Presumably the most common conventional MC questions with one stem, one correct answer and some distracters were used by most authors but many other MC formats exists (e.g. MC questions with Multiple Responses, Complex MC questions, Permutational MC questions, or Liberal MC questions) which could be used, too. Concerning scoring rule the same picture appears. Hence, it is possible that one author used the most common *NC rule* while another author used a different one (e.g. different correction for guessing formulas). This might be one reason why mixed results regarding the similarity of test formats are revealed in previous studies.

Noteworthy is that *NC scoring* for conventional MC questions is an *All-or-Nothing scoring rule* and rewards no partial knowledge. Furthermore, guessing is possible (Ben-Simon et al., 1997; Bereby-Meyer et al., 2002). This is different when MC questions with multiple responses are used, because chances of guessing a question correctly is reduced (Bush, 1999; Foster, 2005), and with *NC scoring* partial knowledge is remunerated.

The article at hand uses MC questions with multiple responses and had an eye on three different scoring rules. Results indicate that CR tests are equal to MC test with multiple responses if *NC scoring* is used. Explanation seems straight forward since the grader of CR tests did not penalize wrong answers and rewarded partial knowledge. Hence, *NC scoring* used the same logic. All other scoring methods either do not reward partial knowledge or penalize guessing. Therefore, these methods are more severe than *NC scoring* or CR tests and cannot be used interchangeably.

Although these two formats (MC tests with multiple responses using *NC scoring* and CR tests) are quite identical it has to be noted that they are not severe enough since the average *logit* of 0 is not met as Figure 3 shows. Furthermore, it is not quite as good as the *WU scoring* to discriminate between students' ability since the majority of students are displayed close together. This discrimination is much better if the *WU scoring* is used (see Figure 2). Therefore, MC tests with multiple responses using *WU scoring* should be the test of choice.

The findings of the study support the widespread usage of MC tests which are superior concerning time and cost reduction (Walstad & Becker, 1994) as well as the greater coverage of the syllabus (Bennett et al., 1990; Walstad & Becker, 1994), and fair grading (Farthing et al., 1998).

With this experiment the authors could demonstrate that not stating the type of MC questions or the scoring rule leads to mixed results which are found in previous studies (e.g. meta-analysis of Rodriguez (2003) or In'nami & Koizumi (2009)).

The authors want to emphasize that a more advanced analysis was used which allows to use smaller data sets. Nevertheless, it has to be mentioned that the sample size was rather small and further research is necessary to replicate results. Furthermore, it has to be noted that grading CR tests is subjective even if scoring criteria are developed (Powell & Gillespie, 1990). In particular, grader always differ in terms of severity/leniency (Engelhard, 1996; Lumley & McNamara, 1993). Hence, follow-up research should use first of all a larger data set and secondly different persons grading the CR exam to overcome these weaknesses. Here, again FACETS is a valuable tool to inspect grading differences. Additionally, there might be other scoring rules not considered in the present study. An extensive literature review to detect additional rules should be a next research step before replicating the study which integrates all scoring rules found.

Literature review also highlighted that findings concerning test formats differ depending on the domain. Therefore, research in other domains would be desirable.

References

Ackerman, T. A., & Smith, P. L. (1988). A Comparison of the Information Provided by Essay, Multiple-Choice, and Free-Response Writing Tests. *Applied Psychological Measurement, 12*(2), 117-128.

American Marketing Association. (2001). First PCM test dates scheduled. *Marketing News, 35*(33).

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., et al. (2000). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition*: Allyn & Bacon.

Bacon, D. R. (2003). Assessing Learning Outcomes: A Comparison of Multiple-Choice and Short-Answer Questions in a Marketing Context. *Jounal of Marketing Education, 25*(1), 31-36.

Becker, W., & Johnston, C. (1999). The relationship between multiple choice and essay response question in assessing economics understanding. *Economic Record, 75*(4), 348-357.

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A Comparative Study of Measure of Partial Knowledge in Multiple-Coice Tests. *Applied Psychological Measurement, 21*(1), 65-88.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a Framework for Constructed-Response Items*. Princeton, New Jerseyo. Document Number)

Bereby-Meyer, Y., Meyer, J., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making, 15*(4), 313.

Berk, R. A. (1986). A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests. *Review of Educational Research, 56*(1), 137-172.

Biggs, J., Kember, D., & Leung, D. Y. P. (2001). The revised two-factor Study Process Questionnaire: R-SPQ-2F. *British Journal of Educational Psychology, 71*(1), 133-149.

Bloom, B. S. (1956). *A Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York: David McKay Company.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bradbard, D. A., Parker, D. F., & Stone, G. L. (2004). An Alternate Multiple-Choice Scoring Procedure in a Macroeconomics Course. *Decision Sciences Journal of Innovative Education, 2*(1), 11-26.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2 ed.). München [u.a.]: Pearson Studium.

Bush, M. (1999). *Alternative Marking Schemes for On-Line Multiple-Choice Tests.* Paper presented at the Seventh Annual Conference on the Teaching of Computing, Belfast, Ireland.

Chan, N., & Kennedy, P. E. (2002). Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple-Choice and "Equivalent" Constructed-Response Exam Questions. *Southern Economic Journal, 68*(4), 957-971.

Crooks, T. J. (1988). The Impact of Classroom Evaluation Practices on Students. *Review of Educational Research, 58*(4), 438-481.

Dressel, P. L., & Schmid, P. (1953). Some Modifications of the Multiple-Choice Item. *Educational and Psychological Measurement, 13*(4), 574-595.

Driscoll, R. (2004). *Westside Test Anxiety Scale*: Westside Psychology.

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Making sense of students' answers to multiple-choice questions. *The Physics Teacher, 40*(3), 174-180.

Eckes, T. (2003). Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse. *Fremdsprachen und Hochschule, 69*, 43-68.

Engelhard, G. (1996). Clarification to “Examining Rater Errors in the Assessment of Written Composition With a Many-Faceted Rasch Model”. *Journal of Educational Measurement, 33*(1), 115-116.

Farthing, D. W., Jones, D. M., & McPhee, D. (1998). *Permutational multiple-choice questions: an objective and efficient alternative to essay-type examination questions.* Paper presented at the ITiCSE '98, Dublin, Ireland

Fisher Jr., W. (1992). Reliability Statistics. *Rasch Measurement Transactions, 6*(3), 238.

Foster, D. (2005). Multiple-Choice Questions Are the Answer. Retrieved 01.09.2009, from www.certmag.com/read.php?in=1171

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research : Innovative quantitative research methods. *Journal of Counseling Psychology, 45*(1), 30-45.

Gibbs, W. J. (1995). An Approach to Designing Computer-Based Evaluation of Student Constructed Responses: Effects on Achievement and Instructional Time. *Journal of Computing in Higher Education, 6*(2), 99-119.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education, 62*(2), 143-157.

Hogan, T. P., & Murphy, G. (2007). Recommendations for Preparing and Scoring Constructed-Response Items: What the Experts Say. *Applied Measurement in Education, 20*(4), 427-441.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*(2), 219-244.

Jennings, S., & Bush, M. (2006). A Comparison of Conventional and Liberal (Free-Choice) Multiple-Choice Tests [Electronic Version]. *Practical Assessment, Research & Evaluation*, *11*, 1-5. Retrieved 22.11.2010, from http://www.pareonline.net/pdf/v11n8.pdf

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of Response Format on Difficulty of SAT-Mathematics Items: It's Not the Strategy. *Jounal of Educational Measurement, 37*(1), 39-57.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2009a). Software: Facets for Many-Facet Rasch Measurement (Version 3.65.0).

Linacre, J. M. (2009b). *A user’s guide to FACETS: Rasch measurement computer program*. Chicagoo. Document Number)

Lukhele, R., Thissen, D., & Wainer, H. (1993). *On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. Program Statistics Research Technical Report No. 93-28*. Princeton, New Jerseyo. Document Number)

Lumley, T., & McNamara, T. F. (1993). *Rater Characteristics and Rater Bias: Implications for Training.* Paper presented at the Language Testing Research Colloquium, Cambridge, UK.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Michigan: University of Michigano. Document Number)

Pollack, J. M., Rock, D. A., & Jenkins, F. (1992). *Advantages and disadvantages of constructed-response item formats in large-scale surveys.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Powell, J. L., & Gillespie, C. (1990). *Assessment: All Tests Are Not Created Equally.* Paper presented at the Annual Meeting of the American Reading Forum, Sarasota, FL.

Reid, F. J. (1976). Scoring Multiple-Choice Exams. *Journal of Economic Education, 8*(1), 55-59.

Rodriguez, M. C. (2003). Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Roediger, H. L., & Marsh, E. J. (2005). The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155-1159.

Rogers, W. T., & Harley, D. (1999). An Empirical Comparison of Three-and Four-Choice Items and Tests: Susceptibility to Testwiseness and Internal Consistency Reliability. *Educational and Psychological Measurement, 59*(2), 234-247.

Rotfeld, H. (1998). Are we teachers or job trainers? *Academy of Marketing Science Quarterly, 2*(August), 2.

Ruch, G. M., & Stoddard, G. D. (1925). Comparative Reliabilities of Five Types of Objective Examinations. *Jounal of Educational Psychology, 16*(2), 89-103.

Schulze, J., & Drolshagen, S. (2006). Format und Durchführung schriftlicher Prüfungen. *GMS Zeitschrift für Medizinische Ausbildung, 23*(3), Doc.44.

Shepard, L. A. (2008). Commentary on the National Mathematics Advisory Panel Recommendations on Assessment. *Educational Reseacher, 37*(9), 602-609.

Traub, R. E., & Fisher, C. W. (1977). On the Equivalence of Contructed-Response and Multiple-Choice Tests. *Applied Psychological measurment, 1*(3), 355-369.

Van den Bergh, H. (1990). On the Construct Validity of Multiple-Choice Items for Reading Comprehension. *Applied Psychological measurment, 14*(1), 1-12.

Ventouras, E., Triantis, D., Tsiakas, P., & Stergiopoulos, C. (2010). Comparison of Examination Methods Based on Multiple-Choice Questions and Constructed-Response Questions Using Personal Computers. *Computers & Education, 54*(2), 455-461.

Wainer, H., & Thissen, D. (1993). Combining multiple choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Walstad, W. B., & Becker, W. E. (1994). Achievement differences on multiple-choice and essay tests in economics. *American Economic Review, Papers and Processing, 84*(2), 193-196.

Ward, W. C. (1982). A Comparison of Free-Response and Multiple-Choice Form of Verbal Aptitude Tests. *Applied Psychological measurment, 6*(1), 1-11.

Weigle, S. C. (1994). Using FACETS To Model Rater Training Effects. *Language Testing, 15*(2), 263-287.

Weiss, B., Gridling, G., Trödhandl, C., & Elmenreich, W. (2006). *Embedded systems exams with true/false questions: A case study* (Research Report No. 5). Vienna: Faculty of Informatics, Vienna University of Technology o. Document Number)

Wright, B., & Linacre, J. M. (2002). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *Journal of Educational Research, 80*(6), 352-358.

Appendix: Examples of the parallel question formats

<table>
<tr><th>CR format</th><th>MC format</th></tr>
<tr><th colspan="2">Question ‘forecasting’</th></tr>
<tr><td>Imagine applying exponential smoothing to forecast your sale. What influence has the smoothing factor ‘α‘ in case it is high?

Example answers:
In the case where α = 1, the forecasted value of the next period is just the same as the original one of the ongoing period.
Values of α close to 1 have less of a smoothing effect, and they give greater weight to recent changes in the data.
Historical time series values are less important for the calculation of the forecast.</td><td>Imagine applying exponential smoothing to forecast your sale. What influence has the smoothing factor ‘α‘ in case it is high? Please select all correct answers.

a) The smoothing effect of time series is rather strong.
b) The adjustment of fluctuations in sale is fast.
c) Consideration of recent values in forecasting is rather low.
d) The forecasted value reacts nervously on fluctuations of the demand.
e) Historical data is weighted strongly.</td></tr>
<tr><th colspan="2">Question ‘research’</th></tr>
<tr><td>You have ordered the following research studies from your Markstrat-Supervisor: ‘Consumer Survey’ and ‘Consumer Panel’. Please explain what you can detect from these studies (especially concerning production, opportunity costs, …).

CONSUMER SURVEY - PURCHASE INTENTIONS
<table>
<tr><th>Firm</th><th>Brand</th><th>Innovs</th><th>Adopters</th><th>Followers</th><th>Total</th></tr>
<tr><td>A</td><td>VAN1</td><td>1.3%</td><td>2.3%</td><td>66.6%</td><td>35.6%</td></tr>
<tr><td>E</td><td>VESI</td><td>35.3%</td><td>52.2%</td><td>5.4%</td><td>25.3%</td></tr>
<tr><td>U</td><td>VUKI</td><td>31.9%</td><td>8.3%</td><td>1.8%</td><td>8.5%</td></tr>
<tr><td>Y</td><td>VYL1</td><td>31.5%</td><td>37.2%</td><td>26.2%</td><td>30.6%</td></tr>
</table>
CONSUMER PANEL - MARKET SHARES BASED ON UNIT SALES
<table>
<tr><th>Firm</th><th>Brand</th><th>Innovs</th><th>Adopters</th><th>Followers</th><th>Total</th></tr>
<tr><td>A</td><td>VAN1</td><td>0.2%</td><td>0.4%</td><td>32.2%</td><td>9.1%</td></tr>
<tr><td>E</td><td>VESI</td><td>60.9%</td><td>72.4%</td><td>17.5%</td><td>55.0%</td></tr>
<tr><td>U</td><td>VUKI</td><td>15.6%</td><td>2.8%</td><td>1.0%</td><td>4.8%</td></tr>
<tr><td>Y</td><td>VYL1</td><td>23.3%</td><td>24.4%</td><td>49.4%</td><td>31.1%</td></tr>
<tr><td colspan="2">Total</td><td>100.0%</td><td>100.0%</td><td>100.0%</td><td>100.0%</td></tr>
<tr><td colspan="2">Total sales(U)</td><td>77,600</td><td>213,800</td><td>110,800</td><td>402,200</td></tr>
<tr><td colspan="2">Total sales(%Total)</td><td>19.3%</td><td>53.2%</td><td>27.5%</td><td>100.0%</td></tr>
</table>
Example answers:
The production plan of company Y was calculated best.
Opportunity costs of company A are very high (119,394 units); opportunity costs of company U are smaller. Nevertheless, both companies need to produce more next period.
Company E could benefit extremely from underproductions of A and U. Hence, they have to be careful when planning next period’s production.</td><td>You have ordered the following research studies from your Markstrat-Supervisor: ‘Consumer Survey’ and ‘Consumer Panel’. Please mark what you can detect from these studies.

CONSUMER SURVEY - PURCHASE INTENTIONS
<table>
<tr><th>Firm</th><th>Brand</th><th>Innovs</th><th>Adopters</th><th>Followers</th><th>Total</th></tr>
<tr><td>A</td><td>VAN1</td><td>1.0%</td><td>0.8%</td><td>26.0%</td><td>19.4%</td></tr>
<tr><td>E</td><td>VESI</td><td>50.3%</td><td>75.0%</td><td>4.3%</td><td>20.9%</td></tr>
<tr><td>U</td><td>VURI</td><td>2.1%</td><td>1.6%</td><td>10.0%</td><td>7.8%</td></tr>
<tr><td></td><td>VUKI</td><td>38.1%</td><td>12.5%</td><td>1.5%</td><td>6.4%</td></tr>
<tr><td>Y</td><td>VYL1</td><td>6.0%</td><td>7.8%</td><td>57.7%</td><td>44.5%</td></tr>
<tr><td></td><td>VYT2</td><td>2.5%</td><td>2.3%</td><td>0.5%</td><td>1.0%</td></tr>
</table>
CONSUMER PANEL - MARKET SHARES BASED ON UNIT SALES
<table>
<tr><th>Firm</th><th>Brand</th><th>Innovs</th><th>Adopters</th><th>Followers</th><th>Total</th></tr>
<tr><td>A</td><td>VAN1</td><td>0.9%</td><td>0.8%</td><td>39.1%</td><td>26,0%</td></tr>
<tr><td>E</td><td>VESI</td><td>62.8%</td><td>83.8%</td><td>5.2%</td><td>30,0%</td></tr>
<tr><td>U</td><td>VUKI</td><td>28.8%</td><td>7.4%</td><td>0.6%</td><td>4,9%</td></tr>
<tr><td></td><td>VURI</td><td>2.6%</td><td>1.6%</td><td>8.5%</td><td>6,3%</td></tr>
<tr><td>Y</td><td>VYL1</td><td>2.7%</td><td>3.8%</td><td>45.6%</td><td>31,3%</td></tr>
<tr><td></td><td>VYT2</td><td>2.2%</td><td>2.6%</td><td>1.0%</td><td>1,5%</td></tr>
<tr><td colspan="2">Total</td><td>100,0%</td><td>100.0%</td><td>100.0%</td><td>100.0%</td></tr>
<tr><td colspan="2">Total sales(U)</td><td>86,100</td><td>226,900</td><td>608,400</td><td>921,400</td></tr>
<tr><td colspan="2">Total sales(%Total)</td><td>9,3%</td><td>24.6%</td><td>66.0%</td><td>100.0%</td></tr>
</table>
a) Company Y could have additionally sold 119,782 units of VYL1.
b) Purchase intention of VYT2 is rather low; hence, brand awareness must be studied to derive correct strategies concerning advertisement etc.
c) VAN1 could benefit from underproductions of other companies.
d) Opportunity costs of company E are rather high because underproduction was 9%.
e) The production plan of company U was calculated best.</td></tr>
</table>