The Impact of Assessment Format on Student and Evaluator Response in Physics Assessment, part I: A Literature Review

James Broberg[1] and Lei Bao[2]

[1] *Department of Physics, Savannah State University, Savannah, Georgia 31404*
[2] *Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA*

(Received August 21, 2012; accepted November 15, 2012)

The purpose of this paper will be to review the current literature studying the impact of assessment format in physics evaluation in order to arrive at a comprehensive picture of the results and effects of different assessment formats on academic performance. This picture shall be obtained by asking two separate questions of each assessment format, and will conclude with a bulleted list of the failures and successes of each assessment format in corresponding to the desired outcome for each question posed. This list shall be used as a guideline for future development of evaluation techniques, as the list of "pros" provide a description of the outcomes desired and the list of "cons" provide a description of the outcomes to be minimized. The two questions that shall organize the discussion and the model that it results in are the following: (a) What is the impact of the assessment format on a student's response?, and (b) what is the impact of the assessment format on the evaluator's response? This paper shall consider the two most prevalent assessment formats, "multiple choice" and "free construction", and subsequent papers will discuss the innovations and alternatives proposed to undercut the dichotomy between multiple choice and free construction while using this paper's model as a guideline for directing progress.

I. INTRODUCTION

The two most common methods used for assessment of academic performance in undergraduate physics classes are the multiple-choice format (hereafter MC) and the free-construction format (hereafter FC). In MC, the student is asked to select one and only one correct option out of a small handful of possibilities with no partial credit given for work shown, while FC gives the student blank paper and a problem and assesses their understanding based on the comprehension shown in the work done to arrive at his answer. The merits and flaws of both approaches have been studied in depth since the first known study of the issue (Starch & Elliot, 1912), with various papers offering different recommendations and suggesting emendations to the formats to provide a more accurate and fair process for evaluating student performance. This paper shall suggest that further clarity can be added to the discussion by distinguishing two distinct questions that need to be addressed separately, and future papers by the present authors will set forth a tentative synthesis incorporating the benefits of both methods while providing a more accurate assessment of student comprehension by employing the use of recent advances in commercially accessible computing technology. Empirical studies for further research into the effectiveness of this technology are currently in the developmental stage.

The decision between MC and FC involves a trade-off between adequacy in assessing student learning, generally regarded as better provided by FC, and efficiency in analyzing and processing the results which MC makes far easier. However, the difference between the formats has been suggested to go beyond this basic trade-off and possibly affect the way students approach the problems, and affect the actual set of skills and aptitudes being evaluated.

This paper shall treat the effect of a problem's format on the student's response separately from the evaluator's response to a problem as being two distinct questions, since the way in which a student approaches a problem or approaches his study is a different fact from the way his performance is evaluated. These two problems are related, because a student's performance (which is affected by the problem's format) in turn partially determines the evaluation of his performance (which is also partially determined by the problem's format). However, if the two aspects are treated separately, one can determine a list of all the factors that would be desirable for both goals, giving a clearer understanding of the ideal "perfect assessment format" to which evaluation should strive to approach, and more concrete steps can be taken to reform the assessment process to approach the ideal and maximize both accuracy and efficiency in student evaluation. The list shall be developed by listing the positive effects of each format structure separately from the negative ones.

II. THE IMPACT OF THE STRUCTURE ON STUDENT RESPONSE

A. The Positive Impact of Multiple Choice Format on Student Response

MC allows and (given time constraints) even forces the student to practice estimation and order-of-magnitude calculation to determine whether a given answer makes

sense or not in order to eliminate obvious wrong answers and guess at the correct one. Many if not most MC format problems do not evaluate estimation and physical intuition, since all the answers are designed to be plausible. However, for evaluating aptitude in this particular skill, MC format is well-suited, and has been used for such in standardized testing (in the 2002 AP chemistry exam, for example).

Because MC evaluates these skills, it requires the students to study for these skills. A 2005 paper notes that the MP format "simplifies a student's learning process considerably (at least that part of it needed to be efficient at exams). The student is focused on the important things from the material" (Raduta & Aubrecht, 2005). This paper follows Hogan (1981) who noted that "[e]vidence collected to date suggests that there are not undesirable side effects, e.g. in terms of students' study habits, resulting from use of choice-type tests" (Hogan, 1981). Whereas FC format permits a student to be sloppy, leave a problem half-finished, or make mistakes and get away with minor penalties in the form of partial credit, MC format requires the student to be able to solve a problem from beginning to end without any mistakes, thereby motivating them to study in order to prepare themselves for perfection.

The authors of the present paper are not convinced, however, as to the extent to which this is unambiguously positive. Is the point of education to prepare students for the exam, or to teach them the material? Just as in research the scientist must apply techniques and models that are known in order to solve problems that are unknown, in most classes offered to students majoring in physics and mathematics it is not sufficient to simply know "important things from the material" – a list of facts or catchwords or highlights – but rather to be able to solve important problems from the material, which requires the application of these catchwords to difficult or conceptually challenging situations, most often situations which the students were not presented with during the lecture itself. It is difficult to write good MC problems that involve the application of familiar material to situations in which creativity and insight would need to be employed, and without the benefit of partial credit grading it is less certain on the part of the evaluator whether the student understands the material or is simply a good test-taker.

The other advantages listed by Raduta and Aubrecht are subject to the same criticism. For example, they state that "the MC questions give the student a finite (a discrete) number of answers, usually four to five. On the other hand, potentially there exist an infinite (a continuous) number of answers from which he usually has to choose the correct one. This is a further simplification which makes the student feel more comfortable… In the long run, this could contribute to the self-confidence and transparency (and sometimes obstinacy and unwillingness to listen) many find characteristic of American culture" (Raduta & Aubrecht, 2005).

Everyone may be sure that the student appreciates the added level of comfort, and reduction of test anxiety in order to evaluate a fuller scope of the student's knowledge is certainly a desired goal which the MC format has been shown to provide (Snow, 1993). But we do not evaluate a fuller scope of the student's knowledge by simplifying the test and making it easier; doing so only results in loss of resolution on the grade-point distributions when the median is raised. If a student is more comfortable because less is required, it can delude him into a false complacency in thinking that he has a greater understanding of the material than he actually does, leaving him unprepared for the more rigorous challenges of more advanced classwork and more difficult FC problems (Chan & Kennedy, 2002).

Secondly, while self-confidence and transparency may (or may not) be values our culture seeks to inculcate, is it the job of a physics exam to mold students into the American cultural model sought for the business world, or is it to teach them comprehension of physics and preparedness for research? It is not clear that the result of this confidence would be positive as a matter of physics education. Certainly, insofar as this increased confidence leads to improved clarity and conciseness of phrasing, it is beneficial, but it is not clearly proven that the MC format gives these results granted all other factors being equal.

A 1998 paper argues that because multiple choice problems are easier and more of them are consequently given in the classroom setting, a wider selection of the course material is evaluated (Saunders & Walstad, 1998a). In rebuttal, one is still stuck with the problem that the problems being given are easier. One can always give a larger number of easier FC problems, but neither scenario will evaluate student comprehension better. As all concepts and methods in physics are co-related and interdependent, a student who performs well on two long and difficult FC problems is likely to understand the basics of all the other concepts as well; one could test this hypothesis by giving students exams with a multitude of easy MC problems and two or three longer FC problems and comparing each student's relative performance, a study that has not yet been done to date.

B. The Positive Impact of Free Construction Format on Student Response

The major positive impact of free construction format on student response is structural fidelity between homework and reality. A comparative study from 2010 argues that FC format prepares students better for professional work in the field by providing a closer simulation of what such work actually entails. "The degree to which examination questions require solving problems similar to those encountered in the actual work situations of a given field" – or structural fidelity – is much higher in FC format than in MC format where the correct answer (unknown in real life) is given to the student alongside distracters (Kuechler & Simkin, 2010).

Another positive impact of free construction format on student response, similar in nature to the point Kuechler and Simkin made, is that it forces students to think through the physics rather than using test-taking skills to eliminate obvious wrong answers and intelligently guess at the correct

one or use estimation. Test-taking skills may hypothetically be valuable skills in some situations, and estimation is certainly a skill every working physicist needs to have. However, in order to show full comprehension of the physics one needs to be able to work through a problem from beginning to end and actually use the correct method for solving a problem rather than simply doing a back-of-the-envelope order-of-magnitude estimation in order to rule out obvious wrong answers. FC format forces a student to be able to work through (or attempt to work through) a problem from beginning to end.

By forcing students to work out a problem from the beginning without showing what the answer might look like, the FC format is more conducive to the application of familiar concepts to unfamiliar situations, encouraging creativity and the extension of the known into the unknown rather than repetition of similar problems already done. This is where real mastery of the concepts is shown, because the student is challenged to think physically and not simply be able to mechanically repeat what was already done in class. Further understanding can be shown by asking the student to carry out derivations of well-known formulae, thereby requiring not just the memorization of sets of equations but actually understanding of what they mean and their relation to other physical principles. One of the authors of the present paper had to work out derivations of various physical laws on exams in almost all of his undergraduate classes; most of them are simple enough to be perfectly feasible for the student to work out alongside other problems in a two-hour time frame.

Occasionally or even frequently a student may not fully comprehend a question or know how to solve it, and in an all-or-nothing MC format be forced guess at the correct answer. Because partial credit is given for work shown in the FC format, the student can work as far as he can before getting stuck and be rewarded proportionally for his efforts. Through this method FC method encourages students to think as hard as possible about the physics in each and every problem, without resorting to guessing or test-taking strategies. This was shown in an empirical study conducted in 1987 which found that students in MC exams committed "a significantly larger number of different error types" than students in FC exams, leading the authors to conclude that "students who have not mastered the task tend to be less consistent in applying their rules of operation for solving procedural tasks when faced with a MC format than with an OE [open-ended] one", for which they attributed different cognitive tasks to each method (Birenbaum & Tatsuoka, 1987).

C. The Negative Impact of Multiple Choice Format on Student Response

The main negative impact of MC format on student response is that it tests a student's aptitude in mastering a multiple choice test, rather than directly testing full comprehension of physics. Raduta and Aubrecht found that students who had managed to correctly answer MC questions were unable to formulate answers to very similar questions in their own words, while students who had managed to correctly answer FC questions had no difficulty in picking similar correct answers in a multiple choice format with lures or distracters (Raduta & Aubrecht, 2005). They explained this result by suggesting that FC questions require a deeper understanding of the underlying physics than MC questions, since there is no prompting or hints from the list of possible correct answers.

The MC format exhibits a structural defect by evaluating a student's test-taking abilities as well as academic performance. A 1986 paper by John Dolly and Kathy Williams, "Using Test-Taking Strategies to Maximize Multiple-Choice Test Scores", showed that an experimental group that had been given a seminar on "testwiseness" with four content-independent test-taking strategies significantly outperformed a control group which had not been given the seminar, showing that cognitive strategies to improve "testwiseness" can not only be performed but even taught (Dolly & Williams, 1986). Further studies (Bush, 2001; Hobson & Ghoshal, 1996) have shown that test-taking strategies improve students' performance on MC exams. By contrast, an ideal exam should strive to evaluate student comprehension and performance in the class, not "testwiseness".

The MC format does not leave room for partial credit for partial comprehension. A student who knows the underlying physics but making a simple calculator error and being forced to guess would receive the same grade as a student who did not know where to even start solving the problem and randomly picked an incorrect answer; a student who guesses after a simple calculator error would receive a lower grade than a clueless student who happens to guess correctly.

Our earlier discussion and critique of Raduta and Aubrecht's study of the comparative confidence and comfort level of students taking MC tests instead of FC tests was skeptical of their argument that simplifying the difficulty level provides a positive benefit to student response. In our support, we cited the paper "Are Multiple-Choice Exams Easier for Economics Students?" by Nixon Chan and Peter Kennedy. Several papers have verified the idea that MC exams are in fact easier than FC exams. Chan and Kennedy pointed out that students can work backwards from the possible answers to the original problems, a luxury they have neither in FC format nor in real life (Chan & Kennedy, 2002). A more recent study by Michael Martinez and Irvin Katz found that the FC format required a greater demand for mental recall, causing the difficulty of the problems to be higher than their MC counterparts (Martinez & Katz, 2010).

FC format requires a greater demand for mental recall because of the inadvertent but ubiquitous cueing effects built into many MC problems. Because MC exams are more difficult to write well than FC exams, the correct answer often stands out among obviously false answers – it is difficult to write good "distracters", as noted by Raduta & Aubrecht (2005). The present authors have used incorrect answers from FC questions written by students in real

classrooms as alternative multiple choice options (as suggested by Cook, 1958), although the number of useful distracters provided by real students is small and the number of ways a student could do a problem incorrectly so large that the possibility of them repeating someone else's error relatively insignificant.

A 1996 study found that difficult items tended to contain negative cueing, directing the student towards the incorrect answer, and vice versa (Donkers et al., 1996), thereby obscuring the test's accuracy in reflecting student comprehension. An earlier paper had shown that cueing tended to favor "poorer" students better than good ones (Harasym et al., 1980). More recent research has presented cueing as an error in calculation: "Despite the fact that MCQs [multiple choice questions] have an advantage concerning objectivity in the grading process and speed in production of results, they also introduce an error in the final formulation of the score. The error is traced to the probability of answering a question by chance or based on an instinctive feeling, which does not enable the ascertainment of the knowledge of the whole background included in the question" (Stergiopoulos et al., 2010). While an instinctive feeling may indicate partial knowledge of the question, a correct answer is evaluated as full knowledge. For all three reasons (positive and negative cueing on different problems, the favoring of students with poorer academic performance, and assessment reflecting "instinctive feeling" rather than full comprehension), cueing provides error in the accuracy of the assessment.

On the other hand, cueing could possibly be viewed as a benefit to the MC format since it can sometimes provide credit for partial knowledge where the FC format does not, by giving the student a hint (in the form of multiple possible answer options) that they can use to get launched on a problem that they would otherwise have no clue how to solve. A 1993 study found a higher rate of skipped problems in the FC format than in the MC format, suggesting that the MC format gave students the hint they needed to get started on the problem, or at least the option to guess (Kingsbury & Houser, 1993). Yet the MC format gives full rather than partial credit for these situations, and a professor administering an FC exam is free to offer hints at his discretion. Since feeding the student information may be regarded as undesirable for assessment purpose, a number of proposals have been set forward for trying to "disguise" the correct answer by including options such as "none of the above" as a possible answer. Unfortunately, at least two papers (Oosterhof & Coats, 1984; Tollefson, 1987) have shown that even when options like "none of the above" are included student performance is still 20-30% higher than on FC exams.

The MC format does not only feed the student information but also misinformation through the form of distracters or lures (incorrect options picked by the student and remembered by the student). Students have an easier time remembering the answers they gave (to which they gave thought and effort of their own) than the solutions they may have been presented with afterwards (to which they did not), and the result can be the retention of wrong answers, an undesired result. "Multiple-choice testing enhances retention of the material tested (the testing effect); however, unlike other tests, multiple-choice can also be detrimental because it exposes students to misinformation in the form of lures. The selection of lures can lead students to acquire false knowledge." (Roediger & Butler, 2008, citing Roediger & Marsh, 2005)

As a result of these negative impacts of the MC format on student response, some authors have been encouraging schools to replace MP evaluation with other formats "to encourage the teaching of higher level cognitive skills" (Frederickson, 1984).

D. The Negative Impact of Free Construction Format on Student Response

As mentioned above, some students will show a fuller exhibition of partial knowledge with the prompting or hints than if they are staring at (and feeling intimated by) a blank page. They may be unwilling to ask for a hint from their professor or unsure as to what to ask. The prompting and cueing provided by MC format can give these students a better chance at demonstrating their knowledge by giving them a hint to start from, or to work backwards from.

Because possible answers are not given to the student in MC format, students in FC format are often unclear as to what the question is asking for or looking for. The MC format gives students a template showing what their final answer should look like. One of the purposes of having a test proctor is to answer questions about what the problem means; however, in the present author's experience as a college instructor, students who do not understand what the problem is looking for are also unaware that they are mistaken and confused, and often do not ask for help. On a basic mathematics diagnostic that one of the authors gives to his students on the first day of class in an introductory calculus-based physics class, for example, students asked to solve for an algebraic variable in an equation with a well-known format (solving for "c" in "$E = mc^2$") will state what they think the variable means physically ("acceleration" was one of the more humorous answers received), rather than the algebraic expression for the variable in terms of the other quantities in the equation ("$\sqrt{(E/m)}$" being the correct answer). The question was not poorly or ambiguously phrased, but the students had come into the class mistaken as to what the word "solve" meant, and an MC format quiz would clarify the format of the answer.

III. Evaluator Response

A. The Positive Impact of Multiple Choice Format on Evaluator Response

One of the most obvious reasons for the use of MC format in the classroom is the ease of grading and administering the exams (Carey, 1997; Frederickson & Collins, 1989; Ramsden, 1988; Scouller, 1988). Their use is more prevalent in larger schools where hand-grading the FC

problems of 500 or more students per school is simply prohibitively expensive (Chan & Kennedy, 2002; Dufresne et al., 2002). For this reason, MC has become the default standard for evaluation for medical exams where a large number of students are being tested (McCoubrie, 2004). Human error can be avoided completely in grading because they can be graded by machines (Holder & Mills, 2001; Kniveton, 1996; Walstad & Becker, 1994, Walstad, 1998). Finally, with a mind toward coordinating standardized exams, Kuechler & Simkin (2004) note that the MC format "helps certification examiners agree on questions to ask a large number of test takers" (Bridgeman, 1991; Bridgeman & Rock, 1993; Holder & Mills, 2001; Snyder, 2004).

Objectivity in grading is a major motivation for the use of MC format (Becker & Johnston, 1999; Thissen & Wainer, 1993; Zeidner, 1987). In the binary MC format, a question is either correct or incorrect, with no middle state. While the advantages of giving partial credit to give better resolution on the assessment of a student's comprehension, the necessarily arbitrary element in partial credit administration by a human, even when following a rubric rigorously, and the imprecise rationale for weighting different parts of the problem differently make it impossible to claim whether partial credit can ever be truly said to be completely "fair". MC format avoids this problem altogether, at the cost of being able to evaluate partial comprehension. MC format also eliminates the inconsistency between students that is unavoidable when a grader is trying to give partial credit in an FC setting (Kniveton, 1996).

The inability to evaluate partial comprehension can be viewed as a positive factor, however. In MC format a student must solve the problem completely correctly or not at all – there is no room for sloppiness or error, and only perfection is expected. This in fact does have structural continuity with professional work in the field, where errors in calculation and solving problems are unacceptable for publication or for homework and exams in some graduate institutions.

Raduta and Aubrecht claim that an advantage to the MC format is that "once one has written some good multiple-choice questions (as measured by appropriate difficulty and discrimination indices, they may be used multiple times (with several classes or in different years), simplifying one's subsequent test-making" (Raduta & Aubrecht, 2005; cf. also Haladyna & Downing, 1989), and also allowing multiple versions of a test to be made to avoid cheating (Kreig & Uyar, 2001; Wesolowsky, 2000). To be sure, this is true, but this claim is just as applicable to good FC questions. The author of the present paper (whose undergraduate and graduate classes were all exclusively FC in format) remembers difficult FC problems from midterm exams appearing on the final exam as a routine practice when a majority in the class showed less than 50% partial comprehension on those problems.

The MC format provides much quicker feedback than the FC format, giving student the opportunity to review the problems they did correctly and incorrectly while it is still fresh in their memory, with positive effect on their education and growth in comprehension (Chan & Kennedy, 2002; Delgado & Prieto, 2003; Epstein, Epstein, & Brosvic, 2001; Epstein & Brosvic, 2002; Kreig & Uyar , 2001).

Finally, the MC format allows evaluator to link reference questions to the test or quiz questions given, so that a student having trouble can study further (Bridgeman & Lewis, 1994). However, this feature can also be adapted to the FC format - it is found for example in the WebAssign program, an online FC educational interface which the authors of the present paper have used for teaching. FC problems given on WebAssign are similar to problems from the textbook but with the numerical values of the quantities in question changed and randomized. The large number of problems that can be given simply by changing the numerical values in the phrasing of the problem gives WebAssign a large bank of distinct problems for students to work for, which has been cited in past literature as a benefit uniquely characteristic of the MC format (Kniveton, 1996; Kuechler & Simkin, 2004).

B. The Positive Impact of Free Construction Format on Evaluator Response

The positive impact of free construction format for the evaluator is that it enables him to see in depth each student's thinking process, bringing to light the misconceptions and problems causing students to have trouble in order to address them directly in the classroom, and painting a detailed picture of each student's status on an individual basis. It enables the instructor to give partial credit to the students, giving a fairer and more accurate assessment of their performance. It gives discretion to the instructor to choose to overlook minor errors which clearly do not affect the student's comprehension (for example, error propagated from previous steps of the problem). FC assesses only the student's comprehension of physics, not their test-taking abilities, thereby providing a more accurate assessment of their performance, and it requires the use of higher cognitive levels than are necessary for the MC format thereby giving the instructor a deeper picture into the student's aptitude.

C. The Negative Impact of Multiple Choice Format on Evaluator Response

The MC format does not evaluate the same cognitive aptitudes and abilities as FC format does, a major problem for the test administrator trying to gain an accurate and fair assessment of student performance. The use of test-taking strategies in MC format has already been discussed above. G. Gage Kingsbury and Ronald Houser argue that MC tests the ability to *recognize* the correct answer while FC tests the ability to *generate* the correct answer: "While multiple choice questions provide an excellent estimate of a test taker's ability to recognize a correct answer to a question,

Table 1. Multiple-Choice Format Advantages and Disadvantages.

	MC Format Pros	MC Format Cons
Student Response	Well-suited for evaluating aptitude in physical intuition and estimation	Lower cognitive levels are required than for FC
	Focuses the student's attention on the most important material	Does not have any allowance for evaluating partial comprehension
	The student feels more comfortable because they are not staring at a blank page	Can leave student with a false sense of complacency that leaves them unprepared
	Because of the student's increased comfort, they are more confident	The MC format evaluates not only physics comprehension but also test-taking strategies
	Allows for assessment of a wider scope of material	Cueing makes it easier
	Requires the student to do all problems perfectly without any mistakes, encouraging better study habits and motivation	Can provide misinformation
		Does not give practice in applying important concepts to non-obvious situations
Evaluator Response	Ease of grading in large classroom settings	Evaluates different cognitive aptitudes and abilities than the FC
	Complete objectivity in grading	Verbal explanations of students who gave correct answers show minimal comprehension of why their answers are even correct
	Evaluator can reference similar questions to the MC's given	
	Evaluator can write multiple versions of the test to avoid cheating	Not a good indicator of performance on equivalent FC exams, performance on subsequent exams, or future academic performance in other classes
	Reusability of good MC problems	
	Quicker feedback	
Technological Integration	If graded on computer, student can receive immediate feedback	

it may be that a test taker's ability to generate a correct answer to a question represents a different and equally important trait to measure" (Kingsbury & Houser, 1993).

There are conflicting results as to whether performance in MC format is a good predictor of performance on FC exams. If it is, then from an evaluator's point of view it is much easier, less expensive, and more efficient to use the MC format. The MC format was widely adopted because many studies did in fact claim that the results are equivalent evaluations of a student's performance, a claim that more recent studies have challenged. A 1991 study funded by the Educational Testing service found that discrepancies between performance on MC and FC questions on the AP computer science exam were statistically insignificant (Bennett et al., 1991), and a 1990 paper in *Applied Psychological Measurement* also found no statistically significant difference (van den Burgh, 1990). In the 1990s this finding was confirmed by Taub, 1993; Thissen & Wainer, 1993; Bennet et al., 1991; Bridgeman, 1991; Bridgeman and Rock, 1993; Walstad & Becker, 1994; Walstad & Kennedy, 1997; and Saunders & Walstad, 1998a. Wainer and Thissen even went so far as to claim that "whatever is… measured by the constructed response section is measured better by the multiple choice section… We have never found any test that is composed of an objectively and a subjectively scored section for which this is not true" (Thissen & Wainer, 1993, cited in Kuechler & Simkin, 2010). A 2006 Physical Review paper comparing the results of MC tests with verbal answers in an introductory physics classroom setting found a discrepancy of only 3%, which is statistically insignificant (Gladding et al., 2006). The most recent claim of the equivalence between MC and FC was a poster gallery presentation presented at the 2011 PERC conference in Omaha by Chandralekha Singh, a professor at the University of Pittsburgh, who more cautiously suggested that "carefully designed multiple-choice assessments can mirror the relative performance on the free-response questions" (Singh & Lin, 2011).

A 1981 paper by Thomas Hogan reviewing past literature on the topic showed that the assessments were "equivalent or nearly equivalent, as defined by their intercorrelation, within the limits of their respective reliabilities", and argued that due to its objectivity in grading that when there were divergences, the MC format was the one to be relied on (Hogan, 1981). The following year, a GRE Board Professional Report by William Ward showed that there was no difference in GRE exams (which

are distributed to thousands of students) between MC and FC formats, and he advocated relying solely on MC format (Ward, 1982).

However, Ward's study was methodologically flawed. The questions he used were too easy to provide a reliable assessment of a student's response to challenging problems involving higher levels of cognition, as he himself noted in his paper (Ward, 1982). A different pool of questions was used for the MC exam as for the FC exam (Ward, 1982), so one cannot make a strict comparison between the two exams. His findings ignore an earlier study conducted over the same GRE material (Vale & Weiss, 1977) which showed that students have to show higher degrees of verbal aptitude to perform at the same level on an FC exam as an MC exam, since the students have to think of the correct answer on their own without prompting (although Vale and Weiss also noted that there could be "more latitude" in the generous grading of a GRE exam giving credit for misspelled words, etc. – cf. also Zeidner, 1987).

A 1996 study conducted by the Educational Testing Institute found that students who did poorly on the MC section of the AP exams and did well on the essay section performed the same in college as students who did well on the MC section but poorly on the essay section (Bridgeman & Morgan, 1996), indicating the equivalence of FC and MC formats. However, tests comparing standardized scores with college performance are problematical because there is no uniformity in the conditions for success in college, with too many different variable factors (environment, quality of education, academic performance of other students at the same institution affecting the grade "curve", etc.) in a college experience to treat them all as equivalent.

Other studies have shown, contradicting the results of the studies already mentioned, that performance on MC exams is a poor predictor of performance on FC exams. The first papers to study the topic were Lorge in 1937 and Lentz in 1938, though due to poor and incomplete publishing of their papers (Rorer, 1965) interest in the topic became dissimulated through a 1946 paper entitled "Response Sets and Test Validity", which argued that the form in which a problem was presented affected the answer given (Cronbach, 1946). Cronbach introduced the concept of a "response set", defined as "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in a different form" (Cronbach, 1946) a tendency caused by personal characteristics on the student which Cronbach called "acquiescence" (Cronbach, 1941, quoted in Hakel, 1998). Another early paper to confirm his findings was an article in *Applied Psychological Measurement* in 1977 (Traub & Fisher, 1977).

However, Cronbach's paper was critiqued and his findings nuanced in a 1965 paper entitled "The Great Response-Style Myth", which argued that Cronbach failed to make a distinction between "response styles" and "response sets", the former being affected by an individual's personality, namely whether they prefer to guess "true" or "false" when they are unsure (Rorer, 1965). Rorer's intuition in "The Great Response-Style Myth" would be expanded into different settings in later work on the problem of middle bias in MC exams (Attali & Bar-Hillel, 2003). Middle bias is a problem in its own right – students guessing on problems they are unsure of are more likely to guess options in the middle rather than at the edges. The data from students guessing creates extra "noise" that obscures the clarity of statistical results (Attali & Bar-Hillel, 2003).

A number of recent papers have found, against the conclusions of the papers mentioned above, that MC exams are poor predictors of performance on FC exams (Carlson et al., 1980; Thissen et al., 1994; Traub & Fisher, 1977; Becker & Johnston, 1999; Hickson & Reed, 2009). Hickson and Reed also showed that MC tests do not accurately reflect performance on subsequent exams in the same course and academic performance in other courses (Hickson & Reed, 2009). Why the disparity in results between different papers? Becker and Johnston critiqued the methodology of earlier papers, showing that a simultaneous equation bias was inherent to the least-squares method of estimating the relationship between the two types of testing (employed by the earlier studies) and that a two-stage least-squares estimation showed no relationship, "implying that these testing forms measure different dimensions of knowledge" (Becker & Johnston, 1999). Dufresne et al. (2002) argued that the so-called "equivalence" is misleading because identical performance does not indicate identical comprehension, since answers on MC questions "more often than not [give] a false indicator of deep conceptual understanding" (Dufresne et al., 2002). "Moderate" relationships were found between performance on MC and FC exams by Kuchler & Simkin (2004) and Bible et al. (2007).

Several studies have even shown that when students are answering MC questions correctly they do not understand why their answers are correct. This was shown in a study of the disparity between students' performance on MC versus FC tests on line graph understanding (Berg and Smith, 1994), and in a comparison between MC problems and verbal explanations of the same problems presented at the 2011 PERC conference (Meltzer, 2011). Kuechler and Simkin mention as "student advantages" that the MC format "[d]oes not require deep understanding of the tested material" (cf. Beard & Senior, 1980; Biggs, 1973; Entwistle & Entwistle, 1992).

D. The Negative Impact of Free Construction Format on Evaluator Response

The problems with objectivity and grading time/manpower have already been discussed under the heading of the positive impact of multiple choice format on evaluator response. Under this heading the only addition to these two points will be a third point, namely the tactical planning of holding exams. Free construction problems are typically longer than multiple choice problems. For example, a final exam for an algebra-based physics class the author taught recently contained 12 multiple choice problems and only 2

Table 2. Free-Construction Format Advantages and Disadvantages.

	FC Format Pros	FC Format Cons
Student Response	Structural fidelity between homework and professional applications of the field	Higher rate of skipped problems due to lack of hints
	Forces students to learn how to work through problems	Does not provide students with template for how the final answer should look, leading to possible confusion
	Challenges the student to think physically and creatively	
	Permits conceptual questions and derivations which force the student to understand the material at a deeper level	
	Encourages the student to try to work through problems where there is only partial comprehension	
Evaluator Response	Allows the grader to see each student's thinking and misconceptions in detail on an individual basis	Takes longer and restricts the scope of material to be tested over
	Allows evaluation of partial comprehension	
	Permits instructor's discretion in awarding full credit where only trivial errors are made	
	Evaluates student performance accurately and fully, not testwiseness	
	Requires student to use higher-level cognitive abilities, allowing them to show their fuller potential	
Technological Integration		Only provided by programs such as WebAssign which anecdotally can be annoying for both student and teacher

free construction problems. A strictly free construction exam would have to contain many fewer problems, thereby restricting the range of material that could be covered, as noted earlier (Becker & Johnston, 1999; Walstad & Robson, 1997; Saunders & Walstad, 1998b; Lukhele et al., 1994). Comments minimizing the severity of this problem were offered earlier.

IV. Additional Impact of Structural Format on Student and Evaluator Response, and Further Cross-Demographic Comparisons

As noted earlier in the paper, there is a third category of the effects of different structural formats on assessment accuracy, caused mainly by the fact that they assess different things rather than giving poorer versus better assessment versus one thing, and these effects cannot be categorized as either unambiguously positive or negative. Also, we must consider comparisons between the two types of formats that do not fit neatly into any of the above categorizations because they are strict comparisons in which the flaws in one method are directly related to the advantages of another.

Raduta and Aubrecht note one such difference between the MC and FC formats in their 2005 paper when they discuss the patterns and types of reasoning that the different formats foster. "This MC system focuses student attention on a discrete-tempered reasoning and by extension may lead students to look at the world as made up of such discrete bits of knowledge, belief, and so on. This discrete-tempered type of reasoning makes the student more efficiently integrated in the real world where this kind of clear, discrete-like-type reasoning structure is much more suitable for being successful in the businesslike environment (where the processes are also discrete-tempered) in which he probably is going to activate… The author has an M.B.A., and has observed this discrete-tempering firsthand" (Raduta & Aubrecht, 2005). This is a double-edged sword, however, as the authors note. The purpose of a physics class is not to prepare future businessmen for their careers, but to teach physics. Furthermore, this simplified space of possible answers is a simplification of reality, and students can be misled to think that the world or that physics is more simplistic than it actually is. "Multiple-choice questions present students with a simplified space (discrete-tempered, one with discrete modes of reasoning, with few alternatives, very clearly formulated in standard ways) corresponding to each question vs. the whole space (a continuous one with continuous modes of reasoning), potentially having an infinite number of answers that the student can formulate to each question. Indeed, the simplified space is a projection of the whole. In the whole space, within the same answer, there exist multiple ways of formulating the same idea, not a standard, optimized, rigid one… [Discrete-tempering] is comfortable also for those people who see the world as rule-bound, and is dangerous to the extent that people view the

discreteness rather than continuity as a characteristic of ideas or pieces of knowledge" (Raduta & Aubrecht, 2005). Raduta and Aubrecht note that when a student switches to a different class that uses a different convention (e.g., different notation), he will be far more confused than if his thinking were not discretized into the nice and pat packages that MC formatting encourages.

Simplification need not be dangerous, however, and a good instructor can use it as an opportunity to illustrate the pervasive method of physics of model-building. The old joke about the farmer whose chicken wouldn't lay eggs turning to a physicist for help and the physicist beginning his explanation of the problem with the statement "Let's start with a spherical chicken" pertains here. We simplify problems by making assumptions, taking limits, ignoring small quantities, and applying models. An instructor could provide a handful of multiple choice questions written in different ways, expressed in different formats, and calculated using different methods in order to illustrate this, and in order to discourage the student from petrifying his thinking into excessive rigidity.

Gender and demographic differences between the MC and FC formats have also been studied, since gender and minority equality are politically and socially fashionable topics today, although no conclusions have been reached from these studies about their reliability in evaluating student performance and giving accurate assessment of cognitive aptitude. Nor has there been any uniformity or consensus concerning how the format affects performance by gender. Beller & Gafni (1996) found no significant gender differences in performance between FC and MC exams in mathematics. A. J. Weaver & Helen Raptis (2001) found no gender differences in introductory atmospheric and oceanic science exams. However, comparisons of AP biology exams and subsequent undergraduate performance have found that multiple choice aptitude correlates to college performance for both males and females, but that performance on the AP essays correlates to college grades for males but not for females (Bridgeman, 1989). The reasons for this remain unknown (the author noted that "further research with large samples is needed for a full understanding of the factors involved").

A handful of authors (Bell & Hay, 1987; Lumsden & Scott, 1987; Bolger & Kellaghan, 1990; Bonner et al., 1994; etc.) have found that males perform better than females on multiple choice tests, and Bridgeman & Lewis (1994) even placed this advantage at a 1/3 standard deviation advantage. This finding has not been subsequently reproduced (Kuechler & Simkin, 2005). It does not follow that the FC is gender-neutral, however, since females tend to outperform males on FC (DeMars, 2000; Simkin & Kuechler, 2005; Simkin & Kuechler, 2010). Other studies have showed no gender gap at all in economics exams (Walstad & Becker, 1994; Greene, 1997; Chan & Kennedy, 2002). More detailed studies of gender difference considering difficulty item have found that in multiple choice exams "Males tended to outperform females on the hardest items; females tended to outperform males on the easiest items" (Bielinski & Davison, 1998).

In summary,

- MC formatting encourages discretized and excessively rigid ways of reasoning, which is both advantageous and dangerous, and is therefore a double-edged sword.
- Contradictory results have been found in looking for differences in gender performance between FC and MC exams. Some studies have found no differences; other studies have found that AP essay response (FC) performance in females does not accurately predict academic performance in college.
- Some studies have shown that males outperform females on MC type exams while females outperform males on FC type exams; other studies have found no significant difference
- One study has shown that males tend to outperform females on the hardest MC items, while females outperform males on the easiest MC items

V. SUMMARY

Having discussed all the positive and negative impacts of both formats on student and evaluator response, it would be useful to summarize our progress in one table. This table shall be used as a guide for future evaluation development, which will seek to achieve as many as possible of the "pros" listed and minimize the "cons".

Our analysis leads us to recommend the continued use of FC format instead of MC while working towards the use of iTest technology for the integration of the desired features of MC into an FC structure. FC format aids the student's progress in learning by providing challenging and creative problems with structural fidelity to real-world work and research, but does not give them any hints or shortcuts. MC format aids the student's educational process by providing a comfortable and streamlined learning environment, one which will be integrated into FC by providing a technological environment; MC harms the student's learning process by providing an artificially easy method of evaluation and even giving misinformation. FC format gives the evaluator aid in his role in education by giving them a fuller insight into the student's response process and greater freedom in recording their assessment, but requires much more time, manpower, and effort, a drawback which can be solved through incorporating computerized grading into the FC structure.

In addition, we found that

- MC formatting encourages discretized and excessively rigid ways of reasoning, which is both advantageous and dangerous, and is therefore a double-edged sword.
- Contradictory results have been found in looking for differences in gender performance between FC and MC exams. Some studies have found no differences; other studies have found that AP essay response (FC)

performance in females does not accurately predict academic performance in college.

- Some studies have shown that males outperform females on MC type exams while females outperform males on FC type exams; other studies have found no significant difference
- One study has shown that males tend to outperform females on the hardest MC items, while females outperform males on the easiest MC items

REFERENCES:

Attali, Y. & and Bar-Hillel, M. (2003). "Guess Where: The Position of Correct Answers in Multiple-Choice Test Items as a Psychometric Variable." Journal of Educational Measurement, 40(2): 109-128.

Aubrecht, G. & Raduta, C. (2005). "Cultural differences in answering physics questions: could they arise from the difference between reasoning expected in answering exam questions in those cultures?" arXiv:physics/0503124v1.

Beard, R. & Senior, I. (1980). Motivating Students. London, UK: Routledge and Kegan Paul.

Becker, W. & Johnston, C. (1999). "The relationship between multiple choice and essay response questions in assessing economics understanding." Economic Record, 75(231): 348-357.

Becker, W. & Walstad, W. (1994). "Achievement differences on multiple-choice and essay tests in economics." The American Economic Review, 84(2): 193-198.

Bell, R. & Hay, J. (1987). "Differences and biases in English language examination formats." British Journal of Educational Psychology, 57, 212-220.

Beller, M. & Gafni, N. (2000). "Can Item Format (Multiple-Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement?" Sex Roles, 42(1/2): 1-21.

Bennett, R., Rock, D., & Wang, M. (1991). "Equivalence of Free-Response and Multiple-Choice Items." Journal of Educational Measurement, 28(1): 77-92.

Berg, C. & Smith, P. (1994). "Assessing Students' Abilities to Construct and Interpret Line Graphs: Disparities Between Multiple-Choice and Free-Response Instruments." Science Education, 76(6):527-554.

Bible, L., Kuechler, W., & Simkin, M. (2007). "How well do multiple-choice tests evaluate students understanding of accounting?" Accounting Education: An International Journal, 17, 55-S68.

Bielinski, J. & Davison, M. (1998). "Gender Differences by Item Difficulty Interactions in Multiple-Choice Mathematics Items." American Educational Research Journal, 35(3): 455-476.

Biggs, J. (1973). "Study behavior and performance in objective and essay formats." Australian Journal of Education, 17, 157-167.

Birenbaum, M. & Tatsuoka, K. (1987). "Open-Ended Versus Multiple-Choice Response Formats – It Does Make a Difference for Diagnostic Purposes." Applied Psychological Measurement, 11(4): 385-395.

Bolger, N. & Kellaghan, T. (1990). "Method of measurement and gender differences in scholastic achievement." Journal of Educational Measurement, 27, 165-174.

Bonner, M. et al. (1994). "Performance versus objective testing and gender." Journal of Educational Measurement, 31(4): 275-293.

Bridgeman, B. (1989). "Comparative Validity of Multiple-Choice and Free-Response Items on the Advanced Placement Examination in Biology." College Board Report No. 89-2; ETS Research Report No. 89-1.

Bridgeman, B. (1991). "Essays and multiple-choice tests as predictors of college freshman GPA." Research in Higher Education, 32(3): 319-331.

Bridgeman, B. & Lewis, C. (1994). "The relationship of essay and multiple-choice scores with grades in college courses." Journal of Educational Measurement, 31(1): 37-50.

Bridgeman, B. & Morgan, R. (1996). "Success in College for Students with Discrepancies Between Performance on Multiple-Choice and Essay Tests". Journal of Educational Psychology, 88(2): 333-340.

Bridgeman, B. & Rock, D. (1993). "Relationships among multiple-choice and open-ended analytical questions." Journal of Educational Measurement, 30, 313-329.

Brosvic, G, Epstein, B., & Epstein, M. (2001). "Immediate feedback during academic testing." Psychological Reports, 88(3): 889-895.

Brosvic, G & Epstein, M. (2002). "Immediate feedback assessment technique: Multiple choice test that behaves like an essay examination." Psychological Reports, 90(1): 226.

Bush, M. (2001). "A multiple choice test that rewards partial knowledge." Journal of Further and Higher Education, 25(2): 157-163.

Butler, A. and Roediger, H. (2008). "Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing." Memory and Cognition, 36(3): 604-616.

Carlson, S., Frederiksen, N., & Ward, W. (1980). "Construct validity of free-response and machine-scorable forms of a test." Journal of Educational Measurement, 17(1): 11-29.

Carey, J. (1997). "Everyone knows that $E = mc^2$ now; who can explain it?" Business Week, 3547, 66-68.

Chan, N. & Kennedy, P. (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.

Coats, P & Oosterhof, A. (1984). "Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple-choice item formats." Applied Psychological Measurement, 8, 287-294.

Collins, A. & Frederickson, J. (1989). "A systems approach to educational testing." Educational Researcher 18:9, 27-32.

Cook, D. (1958). "The Use of Free Response Data in Writing Choice-Type Items." The Journal of Experimental Education, 27(2): 125-133.

Cronbach, L. (1941). "An Experimental comparison of the multiple true-false and multiple-choice tests." Journal of Educationaal Psychology, 32, 533-543.

Cronbach, L. (1946). "Response Sets and Test Validity." Educational and Psychological Measurement, 6, 475-494.

Delgado, A. & Prieto, G. (2003). "The effect of item feedback on multiple-choice test responses." British Journal of Psychology, 94(1): 73-85.

DeMars, C. (2000). "Test stakes and item format interactions." Applied Measurement in Education, 13(1): 55-77.

Dolly, J. & Williams, K. (1986). "Using Test-Taking Strategies to Maximize Multiple-Choice Test Scores." Educational and Psychological Measurement, 46, 619.

Donkers, H., Schuwirth, L., & van der Vleuten, C. (1996). "A closer look at cueing effects in multiple-choice questions." Medical Education, 30, 44-49

Downing, S. & Haladyna, T. (1989). "A taxonomy of multiple-choice writing rules." Applied Measurement in Education, 2(1): 37-50.

Dufresne, R., Gerace, W., & Leonard, W. (2002). "Making sense

of students' answers to multiple-choice questions." The Physics Teacher, 40, 174-180.
Elliott, E. & Starch, D. (1912). "Reliability of the grading of high school work in English." School Review, 20, 442-457.
Entwistle, A. & Entwistle, N. (1992). "Experiences of understanding in revising for degree examinations." Learning and Instruction, 2, 1-22.
Fisher, C. & Traub, R. (1977). "On the Equivalence of Constructed-Response and Multiple-Choice Tests." Applied Psychological Measurement, 1(3): 355-369.
Frederickson, N. (1984). "The Real Test Bias: Influences of Teaching on Testing and Learning." American Psychologist, 39(3): 193-202.
Ghoshal, D. & Hobson, A. (1996). "Flexible Scoring for Multiple Choice Exams." The Physics Teacher, 34(5): 284.
Gladding, G., Scott, M., & Stelzer, T. (2006). "Evaluation multiple-choice exams in large introductory physics classes." Physical Review Special Topics – Physics Education Research, 2, 020102.
Greene, B. (1997). "Verbal abilities, gender, and the introductory economics course: A new look at an old assumption." Journal of Economic Education, 28(1): 13-30.
Hakel, M. (1998). Beyond Multiple Choice: Evaluating Alternatives to Traditional Testing for Selection. Mahwah, NJ: Lawrence Erlbaum Associates.
Harasym, P., Lorscheider, F., & Norris, D. (1980). "Evaluation Student Multiple-Choice Responses: Effects of Coded and Free Formats." Evaluation and the Health Professions, 3(1): 63-84.
Hickson, S. and Reed, W. (2010). "Do Constructed- Response and Multiple-Choice Questions Measure the Same Thing?" Department of Economics Working Paper Series, University of Canterbury (Christchurch, New Zealand).
Hogan, T. (1981). "Relationship between Free-Response and Choice-Type Tests of Achievement: A Review of the Literature." Washington, D.C: National Institute of Education.
Holder, W. & Mills, C. (2001). "Pencils down, computers up: The new CPA exam." Journal of Accountancy, 191(3): 57-60.
Houser, R. & Kingsbury, G. (1993). "A practical examination of the use of free-response questions in computerized adaptive testing." Paper presented to the annual meeting of the American Educational Research Association: Atlanta, Georgia, April 15.
Katz, I. & Martinez, M. (1995). "Cognitive Processing Requirements of Constructed Figural Response and Multiple-Choice Items in Architecture Assessment." Educational Assessment, 3(1): 83-98.
Kennedy, P. & Walstad, W. (1997). "Combining Multiple-Choice and Constructed-Response Test Scores: An Economist's View." Applied Measurement in Education, 10(4): 359-375.
Kniveton, B. (1996). "A correlational analysis of multiple-choice and essay assessment measures." Research in Education, 56, 73-84.
Kreig, R. & Uyar, B. (2001). "Student performance in business and economics statistics: Does exam structure matter?" Journal of Economics and Finance, 25(2): 229-241.
Kuechler, W. and Simkin, M. (2005). "Multiple-Choice Tests and Student Understanding: What is the Connection?" Decision Sciences Journal of Innovative Education, 3(1): 73-97.
Kuechler, W. & Simkin, M. (2010). "Why is Performance on Multiple-Choice Tests and Constructed-Response Tests Not More Closely Related? Theory and an Empirical Test." Decision Sciences Journal of Innovative Education, Volume, 8(1): 55-68.
Lentz, T. (1938). "Acquiescence as a factor in the measurement of personality." Psychological Bulletin, 35, 659.
Lin, S. & Singh, C. (2011). "Correlation between Students' Performance on Free-Response and Multiple-Choice Questions." Gallery Session Poster, Physics Education Research Conference, Omaha, NE.
Lorge, I. (1937). "Gen-Like: Halo or reality?" Psychological Bulletin, 34, 545-546.
Lukhele, K., Thissen, D., & Wainer, H. (1994). "On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests." Journal of Economic Education, 31(3): 234-250.
Lumsden, K. & Scott, A. (1987). "The Economics Student Reexamined: Male-female difference in comprehension." Journal of Economic Education, 18(4): 365-375.
Marsh, E. & Roediger, H. (2005). "The positive and negative consequences of multiple-choice testing." Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 1-7.
McCoubrie, P. (2004). "Improving the fairness of multiple-choice questions: a literature review." Medical Teacher, 26(8): 709-712.
Meltzer, D. (2011). "Time-dependent Interpretation of Correct Responses to Multiple-Choice Questions." Gallery Session Poster, Physics Education Research Conference, Omaha, NE.
Raptis, H. & Weaver, A. (2001). "Gender Differences in Introductory Atmospheric and Oceanic Science Exams: Multiple Choice Versus Constructed Response Questions." Journal of Science Education and Technology, 10(2): 115-126.
Robson, D. & Walstad, W. (1997). "Differential item functioning and male-female differences on multiple-choice tests in economics." Journal of Economic Education, 28, 155-171.
Rorer, L. (1965). "The Great Response-Style Myth." Psychological Bulletin, 63(3): 129-156.
Saunders, P. & Walstad, W. (1998a). "Research on teaching college economics." In Walstad W. & Saunders, P. (eds.), Teaching Undergraduate Economics: A Handbook for Instructors (pp. 141-166). New York: McGraw-Hill.
Saunders, P. & Walstad, W. (1998b). "Using Student and Faculty Evaluations of Teaching to Improve Economics Instruction." In Walstad, W. & Saunders, P. (eds.), Teaching Undergraduate Economics: A Handbook for Instructors (pp. 337-55). New York: McGraw-Hill.
Scouller, K. (1998). "The influence of assessment methods on students' learning approaches: Multiple choice question examinations versus assignment essay." Higher Education, 35, 453-472.
Snow, R. (1993). "Construct Validity and Constructed-Response Tests." In Bennett, R. & Ward, W. (eds.), Construction versus Choice in Cognitive Measurement (pp. 45-60). Hillsdale, NJ: Erlbaum.
Snyder, A. (2003). "The New CPA Exam: Meeting Today's Challenges." Journal of Accountancy, 196(6): 11-120.
Stergiopoulos, C., Triantis, D., Tsiakas, P., & Ventouras, E. (2010). "Comparison of examination methods based on multiple-choice questions and constructed-response questions using personal computers." Computers and Education, 54, 455-461.
Taub, R. (1993). "On the Equivalence of the Traits Assessed by Multiple Choice and Constructed Response Tests." In Bennet, R. and Ward, W. (eds.), Construction versus choice in cognitive measurement. Hillsdale, NJ: Elbaum.
Thissen, D. & Wainer, H. (1993). "Combining multiple-choice and constructed response tests. Toward a Marxist theory of test construction." Applied Measurement in Education, 6(2): 103-118.

Thissen, D., Wainer, H., & Wang, X. (1994). “Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests?” Journal of Educational Measurement, 24(2): 97-118.

Tollefson, N. (1987). “A Comparison of the Item Difficulty and Item Discrimination of Multiple-Choice Items using “None of the Above” and One Correct Response Options.” Educational and Psychological Measurement, 47, 377-383.

Vale, D. & Weiss, D. (1977). “A Comparison of Information Functions of Multiple-Choice and Free-Response Vocabulary Items.” Research Report 77-2, Psychometric Methods Program, Department of Psychology, University of Minnesota, April.

Walstad, W. (1998). “Multiple Choice Tests for the Economics Course.” In William, B. and Phillip, S. (eds.), Teaching Undergraduate Economics: A Handbook for Instructors (pp. 287-304). New York, NY: McGraw-Hill.

Ward, W. (1982). “A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests.” GRE Board Professional Report GREB No. 79-8P; ETS Research Report 81-28, January.

Wesolowsky, G. (2000). “Detecting excessive similarity in answers on multiple choice exams.” Journal of Applied Statistics, 27(2): 909-921.

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

