Advances in Health Sciences Education (2005) 10:133–143
DOI 10.1007/s10459-004-4019-5

The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education

STEVEN M. DOWNING
University of Illinois at Chicago, Department of Medical Education (MC 591), College of Medicine, 808 South Wood Street, Chicago, Il 60612-7309, USA (Phone: +1-312-996-6428; Fax: +1-312-413-2048; E-mail: sdowning@ uic.edu)

Received 10 December 2003; accepted 1 October 2004

Abstract. The purpose of this research was to study the effects of violations of standard multiple-choice item writing principles on test characteristics, student scores, and pass–fail outcomes. Four basic science examinations, administered to year-one and year-two medical students, were randomly selected for study. Test items were classified as either standard or flawed by three independent raters, blinded to all item performance data. Flawed test questions violated one or more standard principles of effective item writing. Thirty-six to sixty-five percent of the items on the four tests were flawed. Flawed items were 0–15 percentage points more difficult than standard items measuring the same construct. Over all four examinations, 646 (53%) students passed the standard items while 575 (47%) passed the flawed items. The median passing rate difference between flawed and standard items was 3.5 percentage points, but ranged from −1 to 35 percentage points. Item flaws had little effect on test score reliability or other psychometric quality indices. Results showed that flawed multiple-choice test items, which violate well established and evidence-based principles of effective item writing, disadvantage some medical students. Item flaws introduce the systematic error of construct-irrelevant variance to assessments, thereby reducing the validity evidence for examinations and penalizing some examinees.

Key words: achievement testing in medical education, construct-irrelevant variance (CIV), flawed test items, item difficulty effects from flawed items, item writing principles, multiple-choice questions (MCQs), pass–fail effects from flawed items, standard test items, written tests

Introduction

Classroom assessment consumes large amounts of instructor time, effort, and resources in medical schools throughout the world. Many current educational measurement textbooks give excellent advice and present thorough instructional materials to assist instructors in preparing effective tests for their students (e.g., Linn and Gronlund, 2000; Nitko, 1996). However, as Mehrens and Lehmann (1991) point out, there are often major deficiencies in examinations prepared by classroom instructors. And, Jozefowicz and others (2002) show that poorly constructed test items are frequently used in medical schools.

The principles of preparing effective objective-test items are well documented (Case and Swanson, 1998; Haladyna, 2004). While test item writing may be as much art as science, there are well established principles, many of which are evidence-based, suggesting what is an effective item form vs. an ineffective item form (Haladyna et al., 2002). Flawed test items result from the violation of one or more of these standard item writing principles.

Several item flaws have been studied empirically for their effect on item and test psychometric characteristics. For example, the use of "all of the above" (AOTA) and "none of the above" (NOTA) as options has been extensively studied with mixed results (Harasym et al., 1998). Variants of the straightforward multiple-choice question (MCQ) stem, such as multiple-true–false or unfocused stems, have been studied and generally found to be detrimental to item performance (Case and Downing, 1989; Downing et al., 1995). Complex item forms, which require selection of combinations of individual options, have been extensively studied and found to be generally detrimental to the psychometric attributes of tests (Albanese, 1993; Dawson-Saunders et al., 1989). The use of negative words in the stem has been evaluated with mixed results concerning difficulty and discrimination of test items (Downing et al., 1991; Tamir, 1993).

A recent review paper (Haladyna et al., 2002) recommends the avoidance of negation in the stem and reports that most educational measurement textbook authors recommend avoiding the AOTA option. The use of the NOTA option has mixed recommendations from textbook authors and the empirical research is also mixed but the current recommendation is to avoid use of the NOTA option, except when used by highly experienced item writers (Crehan and Haladyna, 1991; Frary, 1991).

One small study evaluated the effect of sets of flawed items on the quality indices of an educational achievement test and found that flawed items were generally more difficult and failed more students than comparable standard items (Downing, 2002).

The purpose of this study was to investigate the effect of common multiple-choice item writing flaws on the psychometric characteristics of locally developed medical school achievement examinations used to assess student performance in the basic sciences. Three specific research questions were posed: (1) What is the incidence of item-writing rule violation (flaws) in four achievement tests constructed by medical school basic science faculty? (2) What effect do item flaws have on item difficulty and discrimination and test reliability? (3) What effect do these item flaws have on pass–fail decisions for medical students?

Methods

Four examinations were randomly selected for study from tests routinely administered to first- and second-year medical students, enrolled in pass–fail basic science courses, during the fall semesters of 2001–02 and 2002–03. Two tests were sampled from year-one courses and two tests from year two. Each of the four examinations was from a different basic science discipline. Thus, examination questions were written by different faculty for each test. Student examinees overlap for some examinations.

A *standard* item was operationally defined for this study as any item that did not violate one or more of the 31 principles noted in a recent review article which summarized current educational measurement textbook author recommendations concerning item writing and the empirical research on item flaws (Haladyna et al., 2002). A *flawed* item was operationally defined as an item that violated one or more of these principles.

Test items were classified as either *standard* or *flawed*, and if flawed the exact type of item flaw or flaws contained within the question (including options) was recorded. Three judges, blinded to all item performance data, independently classified each item; there were few disagreements among judges about item classification, and all disagreements were resolved through a consensus process. (Most disagreements concerned multiple flaws within a single test item such that one judge missed the second flaw in the item.)

Example items

Two items serve as examples of the types of item flaws in this study:

1. *It is correct that:*
A. Growth hormone induces production of IGFBP3
B. The predominant insulin-like growth factor binding protein (IGFBP) in human serum is IGFBP3
C. Multiple forms of IGFBP are derived from a single gene
D. All of the above
E. Only A and B are correct

This is an example of an unfocused stem item. The stem does not pose a direct question. The options must each be addressed as "true or false," but the item is scored as a single-best answer question. Further, option D, "all of the above" is not recommended. And, option E is a combination of two other possible answers, making this a "partial-K type" item. Overall, there are three distinct flaws in this question.

2. *Which of the following will NOT occur after therapeutic administration of chlorpheniramine?*
A. Dry mouth
B. Sedation
C. Decrease in gastric acid production
D. Drowsiness
E. All of the above

The second item is an example of a negative-stem question. It requires the student to identify which sign or symptom will not occur. Option *E* (all of the above) is not recommended. This example item has two item flaws.

Examinations studied

For each test selected, three separate scales were scored and item analyzed: the total scale of all items, the set of items classified as standard (standard scale), and the set of items classified as flawed (flawed scale). All items studied were five-option, single-best answer multiple-choice questions (MCQs). All tests were securely administered to student groups in paper-and-pencil format, with students marking answers to questions on optical-scan answer sheets.

A year-one test (Test A) consisted of 72 MCQs and served as the final examination in the discipline. Two items had been eliminated from final scoring by the instructors, due to item content ambiguities. The final six items of this test required examinees to interpret photographs. These six items were eliminated from all scoring and analysis for this study.

A year-two examination (Test B) consisted of 40 questions. This test assessed approximately six weeks of instruction and was administered near the end of a semester. No items were eliminated from scoring or this study.

Test C was a first-year midterm examination in the discipline. This examination consisted of 54 questions. Three items were eliminated from final scoring by the instructors due to various content and psychometric issues and were not included in this study.

Test D was a year-two final examination. A total of 53 total items were used for this study; six items were eliminated from final scoring by the instructors due to content problems.

Scoring for all tests was "right–wrong," with no correction for guessing, and was carried out using a commercial scoring and classical item analysis software package. The tests were administered securely and were timed, but all students had adequate time to answer all examination questions. Absolute passing scores for these examinations were established by faculty content experts, using a modified Nedelsky process which required a judgment about each option of each item (Nedelsky, 1954). Passing standards were established for each test item without access to any performance data and prior to examination administration (Thus, it was possible to compute the number and percentage of students passing and failing subsets of items which had been classified as either standard or flawed).

For each of the three scales evaluated for this study (standard, flawed, and total), item analysis data were computed: raw score means, standard deviations, mean item difficulty, mean point-biserial correlation with the total examination score, Kuder–Richardson 20 reliability (K–R 20), minimum passing score, the number of students passing, and passing rate (the proportion of students who passed).

Results

The descriptive statistics for each total test scale are given in Table I. Tests A and B were easier than Tests C and D and had lower passing scores and higher passing rates than Tests C and D. Internal-consistency reliability of scores ranged from 0.66 to 0.78; mean point-biserial item discrimination indices ranged from 0.18 to 0.21 for these tests. Tests A and B each passed 96% of students; Tests C and D passed 75% and 73%, respectively.

Flawed items comprised 36–65% of the items on these tests. There were a total of 100 (46%) flawed items out of 219 total items on the four examinations.

The use of an unfocused item stem was the most frequent flaw in Tests A and C (Table II). The negative stem flaw ranked first in Tests B and D and was second most frequent in Test A. The "complex matching" form noted in Test A was a unique item form that combined a traditional matching item and a K-type item (complex item format), in which the examinee had to select various combinations of options. "Heterogeneous options" have an option set that is sampled from two or more domains, such that, for example, two

Table I. Descriptive statistics of four tests

	Test A	Test B	Test C	Test D
Number of Items	72	40	54	53
Students	199	179	177	194
Mean item difficulty	0.73	0.77	0.65	0.69
Mean item discrimination	0.19	0.18	0.18	0.21
Reliability	0.78	0.66	0.71	0.76
Minimum pass score				
Raw score	41	22	32	34
Percent	57	55	59	63
Passing rate				
Number students	191	171	133	142
Percent	96	96	75	73

Table II. Frequency of item flaws in four basic science examinations

Type flaw	Test A	Test B	Test C	Test D	Total
Partial K-type	0	5	2	0	7
Complex matching	2	0	0	0	2
Unfocused stem	8	4	13	1	26
Unfocused stem & AOTA	2	0	1	0	3
Unfocused & Negative	0	0	3	3	6
Unfocused stem & NOTA	1	0	2	0	3
Unfocused stem & Partial K	0	1	3	0	4
Unfocused stem, Partial K & AOTA	0	0	1	0	1
Negative stem	6	9	1	7	23
Negative stem & AOTA	0	1	0	0	1
Heterogeneous options	1	0	0	5	6
Heterogeneous options & Partial K-type	0	1	0	1	2
All of the above	7	2	1	0	10
None of the above	0	3	1	2	6
Total flaws	27	26	28	19	100

options may deal with diagnosis while the other three options list laboratory findings. Over all four examinations, the unfocused item stem (in combination with other flaws) and the negative stem (with AOTA) accounted for 67 flawed items (67%). The AOTA option, the NOTA option, and the partial-K type item accounted for an additional 23 flawed items. These five item flaws accounted for 90 of the 100 total flaws.

Slightly more than one-third of the items on Test A were classified as flawed. (Table III). Comparing the psychometric indices for the flawed and standard scales, the only difference is in the mean item difficulty, with the flawed scale 4 percentage points more difficult than the standard scale (95% CI*:* 3–6). Consistent with this finding, the (theoretical) passing rate for the flawed scale is 5 percentage points lower (95% CI: 0–9) than for the standard scale even though the minimum passing score is approximately 5 percentage points lower than the passing score for the standard scale. The correlation between the standard and flawed scales is 0.58 ($p < 0.01$). The correlation corrected for the attenuation of unreliability of both scales is 0.89. (The "correction for attenuation" or disattenuated correlation coefficient estimates the correlation of the two variables if both were perfectly reliable.)

A majority of test questions on Test B were classified as flawed (65%). The standard items are more reliable (when test lengths are equalized) and the mean point-biserial correlation is 4 points higher for the standard vs. the flawed scale. Overall, the standard items are 7 percentage points less difficult than the flawed items (95% CI: 5–9). Four more students (2%) pass the

Table III. Psychometric characteristics of the standard and flawed scales: Tests A, B, C, and D

Test statistics	Test A $N = 199$		Test B $N = 179$		Test C $N = 177$		Test D $N = 194$	
	Standard	Flawed	Standard	Flawed	Standard	Flawed	Standard	Flawed
Items	45	27	14	26	26	28	34	19
K–R 20	0.67	0.63 (0.74)[a]	0.50	0.52 (0.37)[a]	0.55	0.62 (0.60)[a]	0.63	0.55 (0.68)[a]
Mean Diff	0.75 (0.10)	0.71 (0.13)	0.81 (0.13)	0.74 (0.11)	0.73 (0.12)	0.58 (0.14)	0.69 (0.23)	0.69 (0.19)
Mean Discr	0.18	0.20	0.19	0.15	0.15	0.18	0.18	0.18
Mn. Pass Score %	59	54	55	55	58	59	63	63
N Pass	187	178	169	165	151	89	139	142
Pass Percent	94	89	94	92	85	50	72	73

[a] Reliability estimated for the length of the standard scale by Spearman–Brown formula.

standard items as compared to the flawed items. The correlation between the standard and the flawed scales is 0.46 (p <0.0001 This correlation is 0.90 when the effect of unreliability of both scales is removed.

On Test C, the flawed scale is 15 percentage points more difficult than the standard scale (95% CI: 13–17). The flawed scale is more reliable, when test length is adjusted for the shorter standard scale; the flawed scale is also slightly more discriminating than the standard scale. Eighty-five percent (151) of students pass the standard items; 89 students (50%) pass the flawed items. The correlation between the two scales is 0.45 (p <0.0001); corrected for unreliability, the correlation is 0.77.

Test D mean item difficulty and discrimination are equal for both the standard and the flawed scales. The flawed scale is more reliable, when adjusted for the length of the standard scale. Passing scores are the same for both scales; three more students pass the flawed scale as compared to the standard scale. The correlation between the standard and the flawed scales is 0.68 (p < 0.0001) and the disattenuated correlation is unity.

Pass–fail agreement analysis

The agreement in pass–fail outcome status determined by the standard and flawed scales was compared (Table IV). A total of 749 students took all test items. Over all four tests, both the standard and the flawed scales agreed that 544 students (73%) passed and 73 students (10%) failed. Disagreements in assignment of pass–fail status occurred for the 30 students (4%) who passed

Table IV. Pass–fail agreement analysis, all examinations, all students $N = 749$

	Flawed items		
	Fail	Pass	Total
Standard items			
Fail	73	30	103
Pass	102	544	646
Total	175	574	749

the flawed scale but failed the standard and the 102 students (14%) who passed the standard items but failed the flawed questions. Of the 132 (18%) disagreements in assignment of pass–fail status, 102 (77%) showed students passing the standard items but failing the flawed items.

Discussion

This was a non-experimental, descriptive study and consequently generalizations are limited by the circumstances of this study. Insofar as these examinations, test items, and examinees are typically representative of locally developed achievement tests in pre-clinical medical education, the results of this study may generalize to other classroom tests and settings.

In this study, it is important to understand the relationship among item difficulty, item discrimination, score reliability, and passing scores and passing rates. Item difficulty refers to the proportion (%) of students getting the item correct. Item discrimination describes how effectively the test item separates or differentiates between high ability and low ability students – noting that test items that highly discriminate are desirable. All things being equal, highly discriminating items tend to produce high score reliability. Because items in this study had each been assigned a passing score value (by the Nedelsky absolute standard setting method) it was possible to calculate passing scores (the score needed to pass the test) and passing rates (the percentage of students who pass) for the two subscales of interest – the standard and the flawed subscales. It is important to note that the passing score and the passing rate are inversely related; that is, high passing scores tend to produce lower passing rates.

There was a high frequency of flawed items in the tests studied. This is an important finding, although not completely unexpected (Downing, 2002; Jozefowicz et al., 2002; Mehrens and Lehmann, 1991). Classroom assessments in medical school settings are not immune to poorly crafted test items. The item flaws studied were non-subtle, obvious violations of the well established principles of effective multiple-choice item writing. Unfocused item stems, negatively worded stems, use of the AOTA and the NOTA option, and oddly designed complex item formats account for the majority of item flaws found in this study.

Flawed item formats were more difficult than standard, non-flawed item formats for students in three of four examinations studied. These mixed results showed that flawed item formats were 0–15 percentage points more difficult than questions posed in a standard form. This finding is somewhat surprising, given that examinees in this study are medical students, highly experienced in taking MCQ examinations and presumably very testwise.

Passing rates (the proportion of students meeting or exceeding the passing score) tended to be negatively impacted by flawed items. Poorly crafted, flawed test questions tended to present more of a passing challenge for students.

The agreement between pass–fail outcome assigned by the standard and the flawed scales shows that 102 of 749 students (14%) pass the standard items but fail the flawed items, while only 30 students (4%) pass the flawed items and fail the standard items. (These data must be interpreted cautiously, since the scales differ in length and reliability and the passing scores also differ for some of the scales.) Since no test or test scale can be perfectly reliable, there will always be some error in classifying students as passing or failing. The 102 students (of 749) classified as passing the standard items while failing the flawed items are of great concern. Some of these misclassifications are due to random measurement error (unreliability), but some proportion is also due to the systematic error introduced by flawed items, given the results of this research.

One can conclude that some students – perhaps as high as 10–15% of students tested – were incorrectly classified as failed when they should have been classified as passed, due solely to flawed item formats and the ineptitude of test item writers. A false negative rate this high seems unreasonable, given the relative ease and low costs associated with re-writing flawed questions into a form that would adhere to the standard, evidence-based principles of effective item writing. Clearly, this high misclassification rate impacts the consequential validity evidence for the tests in a negative manner (Messick, 1989).

The effect of flawed item forms on score reliability is mixed; in three of the four tests studied, the estimated score reliability was actually higher for the flawed subscale compared to the standard subscale. The nature of the item format flaws studied contributes systematic error to the measurement, not random error; only random errors of measurement are estimated by the internal consistency score reliability. Thus, it is not surprising that the score reliability shows little relationship to item flaws.

The additional test difficulty introduced into the measure by poorly crafted and flawed item formats is an example of construct-irrelevant variance. Messick (1989, p. 34) defines construct-irrelevant variance (CIV) as "…excess reliable variance that is irrelevant to the interpreted construct." The excess difficulty and tendency toward lower passing rates for flawed vs. standard items meets Messick's definition of CIV perfectly.

The disattenuated correlation estimates the true score correlation between both scales and should be near unity, since both the standard and the flawed scales measure the identical construct. The square of the disattenuated correlation between the standard and the flawed scale may be considered an index of the CIV contributed by flawed items to these assessments, such that 1-R_t^2 is an estimate of the CIV error variance contained in the assessment. In this research, flawed test questions contributed from 0 to 41% (median = 20%) of CIV error variance to the total variance of test scores.

Further study of the effects of using flawed item forms and their contribution to CIV should be undertaken. Studies designed to determine the effect of testwiseness and student aptitude, as measured by some reliable external criterion, are needed to more completely understand how students interact with flawed test questions.

The results of this study suggest that efforts to teach faculty the principles of effective objective-test item writing should be increased. The good news is that these faculty development efforts can concentrate on eliminating the five most common errors found in this study and thereby eliminate nearly all flawed items from tests. Other methods to reduce or eliminate flawed items or other non-standard item formats from locally developed achievement tests should be developed and implemented, especially in settings where the stakes associated with achievement measurement are moderate to high.

Acknowledgements

The author is grateful to Profs. Thomas M. Haladyna and Mark Albanese for their critical review of this manuscript. Also, the author gratefully acknowledges the assistance of Ms. Janeille Egbert and Mr. Robert Smith, University of Illinois at Chicago, Department of Medical Education, Testing Service.

References

Albanese, M. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practices* **12**: 28–33.

Case, S.M. & Downing, S.M. (1989). Performance of various multiple-choice item types on medical specialty examinations: Types A,B,C, K, and X. *Proceedings of the Twenty-Eighth Annual Conference on Research in Medical Education*, pp. 167–172.

Case, S.M. & Swanson, D.B. (1998). *Constructing Written Test Questions for the Basic and Clinical Sciences,* 2nd edn. Philadelphia, PA: National Board of Medical Examiners.

Crehan K.D. & Haladyna T.M. (1991). The validity of two item-writing rules. *Journal of Experimental Education* **59**: 183–192.

Dawson-Saunders, B., Nungester, R.J. & Downing, S.M. (1989). A comparison of single best answer multiple-choice items (A-type) and complex multiple-choice items (K-type). *Proceedings of the Twenty-Eighth Annual Conference on Research in Medical Education*, pp. 161–166.

Downing, S.M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item writing principles make any difference? *Academic Medicine* **77**: s103–104.

Downing, S.M., Baranowski, R.A., Grosso, L.J. & Norcini, J.J. (1995). Item type and cognitive ability measured: The validity evidence for multiple true–false items in medical specialty certification. *Applied Measurement in Education* **8**: 89–199.

Downing, S.M., Dawson-Saunders, B., Case, S.M. & Powell, R.D. (April, 1991). The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II characteristics. A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Frary, R.B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education* **4**: 115–124.

Haladyna, T.M. (2004). *Developing and Validating Multiple-choice Test Items*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T.M., Downing, S.M. & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines. *Applied Measurement in Education* **15**: 309–334.

Harasym, P.H., Leong, E.J., Violato, C., Brant, R. & Lorscheider, F.F. (1998). Cuing effect of "all of the above" on the reliability and validity of multiple-choice test items. *Evaluation and the Health Profession* **21**: 120–133.

Jozefowicz, R.F., Koeppen, B.M., Case, S., Galbraith, R., Swanson, D. & Glew, H. (2002). The quality of in-house medical school examinations. *Academic Medicine* **77**: 156–161.

Linn, R.L. & Gronlund, N.E. (2000). *Measurement and Assessment in Teaching,* 8th edn. Upper Saddle River, NJ: Prentice-Hall.

Mehrens, W.A. & Lehmann, I.J. (1991). *Measurement and Evaluation in Education and Psychology*. New York: Harcourt Brace.

Messick S. (1989). Validity. In R.L. Linn (ed.), *Educational Measurement* (3rd edn.), pp. 13–104. New York: American Council on Education and Macmillan.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement* **14**: 181–201.

Nitko, A.J. (1996). *Educational Assessment of Students*. Englewood Cliffs, NJ: Merrill.

Tamir, P. (1993). Positive and negative multiple choice items: How difficult are they? *Studies in Educational Evaluation* **19**: 311–32