The Canadian Journal for the Scholarship of Teaching and Learning

Volume 2 | Issue 2 Article 4

12-23-2011

Examination of the Quality of Multiple-choice Items on Classroom Tests

David DiBattista
Brock University, david.dibattista@brocku.ca

Laura Kurzawa
Brock University, laura_kurz@hotmail.com

Follow this and additional works at: http://ir.lib.uwo.ca/cjsotl_rcacea
http://dx.doi.org/10.5206/cjsotl-rcacea.2011.2.4

Recommended Citation
DiBattista, David and Kurzawa, Laura (2011) "Examination of the Quality of Multiple-choice Items on Classroom Tests," *The Canadian Journal for the Scholarship of Teaching and Learning*: Vol. 2: Iss. 2, Article 4.
DOI: http://dx.doi.org/10.5206/cjsotl-rcacea.2011.2.4
Available at: http://ir.lib.uwo.ca/cjsotl_rcacea/vol2/iss2/4

Examination of the Quality of Multiple-choice Items on Classroom Tests

Abstract
Because multiple-choice testing is so widespread in higher education, we assessed the quality of items used on classroom tests by carrying out a statistical item analysis. We examined undergraduates' responses to 1198 multiple-choice items on sixteen classroom tests in various disciplines. The mean item discrimination coefficient was +0.25, with more than 30% of items having unsatisfactory coefficients less than +0.20. Of the 3819 distractors, 45% were flawed either because less than 5% of examinees selected them or because their selection was positively rather than negatively correlated with test scores. In three tests, more than 40% of the items had an unsatisfactory discrimination coefficient, and in six tests, more than half of the distractors were flawed. Discriminatory power suffered dramatically when the selection of one or more distractors was positively correlated with test scores, but it was only minimally affected by the presence of distractors that were selected by less than 5% of examinees. Our findings indicate that there is considerable room for improvement in the quality of many multiple-choice tests. We suggest that instructors consider improving the quality of their multiple-choice tests by conducting an item analysis and by modifying distractors that impair the discriminatory power of items.

Étant donné que les examens à choix multiple sont tellement généralisés dans l'enseignement supérieur, nous avons effectué une analyse statistique des items utilisés dans les examens en classe afin d'en évaluer la qualité. Nous avons analysé les réponses des étudiants de premier cycle à 1198 questions à choix multiples dans 16 examens effectués en classe dans diverses disciplines. Le coefficient moyen de discrimination de l'item était +0.25. Plus de 30 % des items avaient des coefficients insatisfaisants inférieurs à + 0.20. Sur les 3819 distracteurs, 45 % étaient imparfaits parce que moins de 5 % des étudiants les ont choisis ou à cause d'une corrélation négative plutôt que positive avec les résultats des examens. Dans trois examens, le coefficient de discrimination de plus de 40 % des items était insatisfaisant et dans six examens, plus de la moitié des distracteurs était imparfaits. Le pouvoir de discrimination était considérablement affecté en cas de corrélation positive entre un distracteur ou plus et les résultatsde l'examen, mais la présence de distracteurs choisis par moins de 5 % des étudiants avait une influence minime sur ce pouvoir. Nos résultats indiquent que les examens à choix multiple peuvent être considérablement améliorés. Nous suggérons que les enseignants procèdent à une analyse des items et modifient les distracteurs qui compromettent le pouvoir de discrimination des items.

Keywords
multiple choice, assessment, classroom testing, item discrimination, distractor analysis

Cover Page Footnote
We thank the Brock University Centre for Teaching, Learning and Educational Technologies for a grant to support this research, and we thank the instructors who so kindly participated in this research study.

This research paper/rapport de recherche is available in The Canadian Journal for the Scholarship of Teaching and Learning:
http://ir.lib.uwo.ca/cjsotl_rcacea/vol2/iss2/4

Multiple-choice (MC) items are widely used on classroom tests in colleges and universities and they often account for a substantial portion of a student's course grade (Mavis, Cole, & Hoppe, 2001; McDougall, 1997). A typical MC item consists of a question, referred to as the stem, and a set of two or more options that consist of possible answers to the question. The student's task is to select the one option that provides the best answer to the question posed. The best answer is referred to as the keyed option and the remaining options are called distractors. For instructors, a distinct advantage of using MC items on classroom tests is that grading tends to be quick and easy, especially when students indicate their answers on an optically scanned MC response sheet, such as the widely used Scantron® form. Ease of grading can make MC testing particularly appealing to instructors who teach courses with large enrolments. Another important advantage is that a well-constructed MC test can yield test scores at least as reliable as those produced by a constructed-response test, while also allowing for broader coverage of the topics covered in a course (Bacon, 2003).

Despite these advantages, MC testing is often criticized. Some authors have pointed out that MC items focus on what students can remember and do not assess the extent to which they can understand, apply and analyze course-related information (Walsh & Seldomridge, 2006). However, it is clear that thoughtfully written MC items can serve to assess higher-level cognitive processes, although creating such items does require more skill than writing memory-based items (Buckles & Siegfried, 2006; Palmer & Devitt, 2007). Another criticism is that the format of MC items lets students guess even when they have no substantive knowledge of the topic under consideration (Biggs, 1999). However, Downing (2003) points out that blind guessing is quite uncommon on well-written classroom tests and informed guessing, which is based on a critical consideration of the question and the available options, provides a valid measure of student achievement.

Whatever one's opinion about its merits, MC testing is very widely used to assess student achievement in postsecondary classrooms. Furthermore, financial constraints currently faced by many educational institutions will likely lead to increases in class size (Schrecker, 2009), which may in turn lead to increased use of MC testing in the future. Given the extensive use of MC testing in postsecondary settings, it seems prudent to look carefully at the quality of the MC items on classroom tests, and this was the primary purpose of the research that will be reported here. Of course, there are many factors to consider when evaluating the quality of MC items. For example, one might examine the extent to which items conform to widely accepted item-writing guidelines, such as putting the central idea of the question into the stem and avoiding the use of negation whenever possible (Haladyna, Downing, & Rodriguez, 2002). Deviating from the guidelines can be problematic because it can detract from the quality of individual items and of the test as a whole (Downing, 2005; Tarrant & Ware, 2008). As it happens, failure to conform to the guidelines is widespread in the MC items found both on tests in postsecondary classrooms (Jozefowicz et al., 2002; Tarrant, Knierim, Hayes, & Ware, 2006) and in publisher-supplied test banks (Hansen & Dexter, 1997; Masters et al., 2001).

Another way to examine the quality of MC items involves analyzing the responses that examinees make, and this is the approach used in the research presented here. Specifically, we analyzed instructor-designed tests administered to students in undergraduate university

classrooms and focused on three key characteristics of individual MC items: difficulty, discriminatory power, and effectiveness of the distractors. An overview of these characteristics follows.

When students have taken a MC test, the difficulty index of an item is the proportion of examinees who selected the keyed option. The difficulty index, symbolized as p, can range from 0 (no one selected the keyed option) to 1.00 (everyone selected it). Naturally, overall test scores tend to be higher when the items on a test have higher p values, and vice versa.

A major determinant of the quality of a MC item is its discriminatory power (Ebel, 1975), which reflects the extent to which more knowledgeable students are more likely than less knowledgeable students to select the keyed option. The discriminatory power of a MC item can be measured by computing its discrimination coefficient, which is the correlation between examinees' overall test scores and the scores that they have obtained on the item under consideration (i.e., 1 if they selected the keyed option, and 0 otherwise). The discrimination coefficient, symbolized as r_{PBS}, is a point-biserial correlation that is mathematically equivalent to the more familiar Pearson r and is interpreted in much the same way. Thus, for a MC item to function effectively, its discrimination coefficient must be a positive value, which indicates that examinees with higher test scores performed better on the item than did those with lower scores. In addition, most authors suggest that the discrimination coefficient should be at least +0.20 (Ding & Beichner, 2009; Su, Osisek, Montgomery, & Pellar, 2009; Thorndike, 2005), although some place this benchmark either somewhat lower (Kehoe, 1995: +0.15) or higher (Considine, Botti, & Thomas, 2005: +0.25). When an item's discrimination coefficient is positive but small, it is not discriminating sufficiently between the higher- and lower-scoring examinees to contribute to the overall quality of the test. Even more problematic are items that function so poorly that they have a negative discrimination coefficient, perhaps because the wording is unclear or because two options rather than one are correct (Reid, 1970). Such items detract from the overall quality of a test because examinees with lower test scores are selecting the keyed option more often than those with higher scores.

When an item's difficulty index is either very low or very high, its discriminatory power tends to suffer. For instance, consider the extreme case in which all examinees answer an item correctly (or incorrectly) – that is, the difficulty index is equal to one (or zero). Under these circumstances, item scores are uncorrelated with total test scores, and the item will have no discriminatory power at all. Generally speaking, MC items that are either very difficult ($p<0.30$) or very easy ($p>0.90$) tend to do a rather poor job of discriminating between higher and lower achievers (Ebel & Frisbie, 1991).

The discriminatory power of a MC item depends heavily on the quality of its distractors. An effective distractor will look plausible to less knowledgeable students and lure them away from the keyed option, but it will not entice students who are well-informed about the topic under consideration. Writing effective distractors can be challenging, but helpful guidelines that can make the process easier are readily available (Haladyna, 2004; McDonald, 2007). Suggestions include, for example, using errors that are commonly made by students, using true statements that do not correctly answer the question posed in the stem, and avoiding the use of all-of-the-above as an option.

From a functional perspective, a distractor must meet two criteria in order to be effective. First, at least some examinees must select it. If they do not, then the distractor is not luring anyone

away from the keyed option, and it cannot contribute to the item's discriminatory power. Haladyna & Downing (1993) have suggested that at least 5% of examinees should select each of an item's distractors, and this value is a common benchmark for distractor functionality (Tarrant, Ware, & Mohammed, 2009; Ware & Vik, 2009). The second criterion relates to a distractor's ability to discriminate between stronger and weaker examinees. Recall that for a MC item to have good discriminatory power, examinees with higher test scores must select the keyed option more often than those with lower scores. For a distractor to be effective, the opposite must be true – that is, examinees with higher test scores must select the distractor *less* often than those with lower scores. When this happens, examinees' selection of the distractor will be *negatively* correlated with total test scores. Conversely, a distractor that has either a positive or a zero correlation with total test scores is not functioning properly and detracts from an item's overall quality.

It is clear then that statistical techniques are available for assessing the quality of MC items used on classroom tests. As it happens however, research on item quality has mostly involved large-scale standardized tests, with relatively little published research focusing on classroom assessment (Stiggins & Bridgeford, 1985). The available research suggests that the mean item discrimination coefficient for classroom tests most often lies somewhere between 0.20 and 0.30, with a substantial proportion of items being less than satisfactory discriminators (Martinez, Moreno, Martin, & Trigo, 2009; Phipps & Brackbill, 2009; Tarrant et al., 2009). For instance, Oppenheim (2002) looked at a business law exam that consisted of 66 MC items taken from a test bank developed by a task force of the Academy of Legal Studies in Business. When a sample of 41 students took this test, the mean ± SD item discrimination coefficient was 0.24 ± 0.17. It is noteworthy that despite the great care that went into the development of these test items, more than one-third had discrimination coefficients less than 0.20.

Several studies have looked at how the difficulty and discriminatory power of MC items change when dysfunctional distractors are either replaced or deleted (e.g., Cizek & O'Day, 1994), but few have looked at the quality of distractors in MC tests specifically designed for classroom use. In a recent study of four year-end medical school tests containing 389 MC items, Ware & Vik (2009) considered any distractor selected by at least 5% of examinees to be functional. Using this lenient definition, which ignores whether distractor selection is negatively correlated with test scores, they found only 36% of 1557 distractors to be functional. Tarrant et al. (2009) more appropriately defined a functional distractor as one that was selected by at least 5% of examinees and was also negatively correlated with test scores. By this definition, only 52% of 1542 distractors on seven nursing tests functioned properly. Furthermore, 12% of items had no functional distractors, and 35% had only one. It must be noted that Tarrant et al. specifically excluded from their analysis tests with a reliability of less than 0.70; if tests with lower reliability had been included, it is likely that the percentage of functional distractors would have been even lower. In summary, the available evidence suggests that many, if not most, of the distractors that are used on classroom tests function quite poorly.

In this report, we present findings from a study of the MC items used on classroom tests taken by undergraduate students at a mid-sized Canadian university. In contrast to earlier studies, which have involved looking at a small number of tests in a single discipline, we looked at 16 tests in a variety of disciplines. We focused our attention on issues relating to item difficulty, item discrimination, and the effectiveness of distractors. We also present the results of a survey of university instructors on issues related to testing. Anecdotal evidence suggests that MC testing is most common in larger classes and in the lower years of the curriculum, and that it is rarely used

in humanities courses (Cirino-Gerena, 1981). Furthermore, it is well known in the postsecondary community that most instructors have little or no formal training in pedagogy (McDougall, 1997; Ravenscroft, Rebele, St. Pierre, & Wilson, 2008). The goal of the survey was to learn more about instructors' background in issues relating to testing and about the extent to which they use MC items on classroom tests.

Method

We initially selected a total of 240 different undergraduate courses being offered during the fall, winter and spring semesters at a mid-size university in Ontario, Canada. In selecting courses, we randomly picked 12 courses at each of the four year levels within each of five faculties (Applied Health Sciences, Business, Humanities, Mathematics & Science, and Social Sciences). We did not include in the sample any courses that would not normally be expected to involve in-class assessments (e.g., honours thesis, directed reading). We made sure that all courses were taught by different instructors, and we sent the 240 instructors a letter inviting them to complete a survey. If they did not respond, we sent one follow-up letter three weeks later. In the survey, we asked instructors to indicate whether they used MC items on classroom tests in the selected course, and if so, what percentage of the total course marks were derived from MC items. We also asked them to indicate how they had learned to construct tests and to interpret test results. We received completed surveys from 116 instructors.

After collecting the survey data, we contacted the 59 instructors who had indicated that they used MC items in the selected course. We invited these instructors to provide us with the MC response sheets submitted by their students for a summative test given in the course, and we specifically asked that instructors choose a test containing as many MC items as possible. A total of 38 instructors gave us MC response sheets, which were submitted to the university's Information Technology Services Department for optical scanning and scoring (1 point for correct responses and 0 for incorrect responses). The resulting computer-generated report provided a variety of descriptive statistics for the test, and for each item, it showed the difficulty index, discrimination coefficient, and number of examinees who selected each distractor. The report also sorted examinees into quartiles based on their overall test scores (1=lowest and 4=highest), and for each item it computed the frequency with which examinees in each quartile selected each distractor. The discrimination correlation for each distractor was then determined by computing the correlation (Pearson r) between the distractor's four selection frequencies and their respective quartiles (i.e., 1, 2, 3, or 4). For a properly functioning distractor, this discrimination correlation would be negative. For purposes of statistical analysis, we considered a properly functioning distractor to be one that had a negative discrimination correlation and was selected by at least 5% of examinees.

To ensure the trustworthiness of our analyses (Bodner, 1980), we present here only the data for the 16 submitted tests for which there were at least 24 MC items and 100 examinees. We conducted statistical analyses using release 18.0.0 of IBM SPSS Statistics (SPSS Inc., Chicago, Illinois). The university's Research Ethics Board reviewed and approved all procedures.

Results

Survey Results

Of the 116 instructors completing the survey, one asked us not to use information about the year level and faculty of his/her course. The data set included roughly equal numbers of courses at each year level (maximum: Year 1=31; minimum: Year 4=27), with somewhat more variability across faculties (maximum: Social Sciences=30; minimum: Math & Science=17). According to data from the university's Registrar, the mean class size was 95.15 ± 108.51; the distribution of class sizes was positively skewed, with a median of 55.0, a low of 2 and high of 497.

Of the instructors who responded, 51% reported using MC items on tests in the course under consideration. MC usage was widespread in the Applied Health Science (76% of courses), Business (62%), Social Sciences (55%), and Mathematics & Science (41%). In contrast, only one course in Humanities (5%) used MC items. As Figure 1 indicates, MC items were used more often in lower-year than in upper-year courses, χ^2 (3) = 15.10, $p < 0.01$, $\varphi = 0.36$. When used, MC items accounted for a mean of 30.74 ± 23.66 percent of total course marks. In 22% of the 59 courses using MC items, they accounted for more than half of total marks, and in two courses, they accounted for 90% of marks.

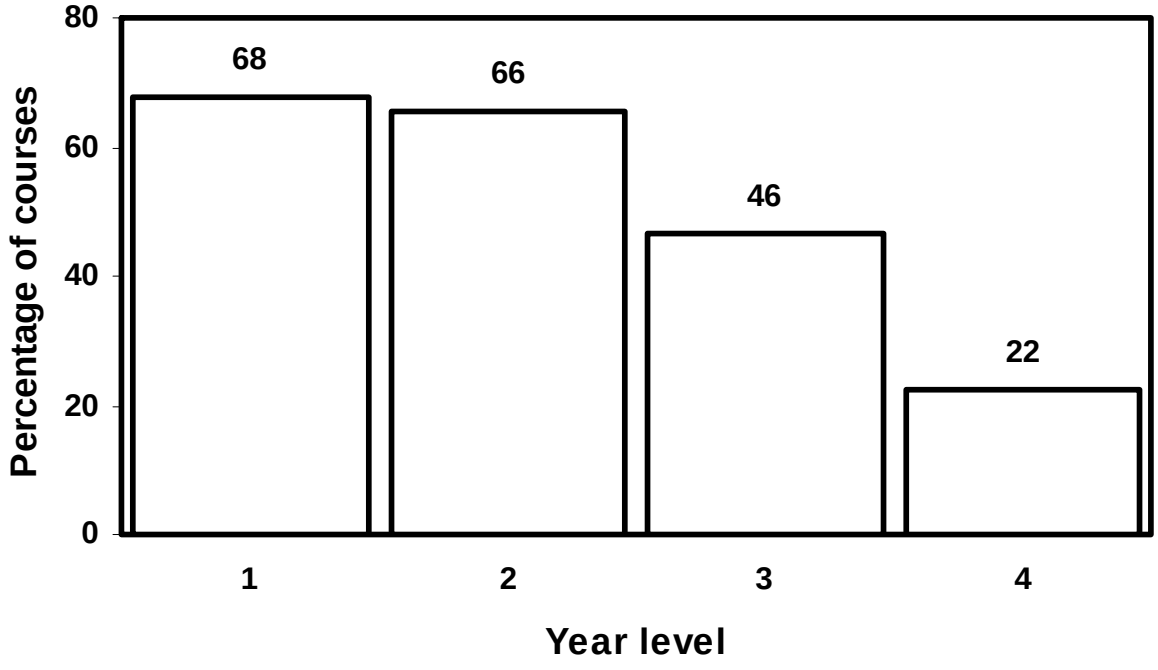

Figure 1. Percentage of undergraduate courses using MC items on tests.

Class size was strongly and inversely associated with year level, r (113) = −0.59, p < 0.001. Not surprisingly therefore, enrolment was higher in courses that used MC items (*Med* = 95.0) than in those that did not (*Med* = 32.0), Mann-Whitney z= 5.30, p < 0.001. When MC items were used, the percentage of marks that they accounted for was not statistically significantly associated with the year level of tests, r (57) = −0.23, n.s. However, the percentage of course marks was directly related to class size, r (57) = 0.34, p < 0.01, and this association remained statistically significant even when year level was partialled out, r (56) = 0.27, p < 0.05.

Only 33% of instructors reported having taken formal courses dealing with test construction and the interpretation of test results. Instructors who had never taken such courses said that their knowledge of testing was based primarily on factors such as personal experience (94%), interactions with peers (72%), trial and error (68%), books (37%), and workshops (23%).

Analysis of MC Items

Table 1 summarizes the characteristics of the 16 tests that were examined, most of which were at the first- and second-year level. In all, there were 1198 MC items. The number of items on tests ranged from 24 to 211, and the number of examinees ranged from 109 to 547. Mean test scores, which are reported as percentages, were highly variable, ranging from 45.2 to 73.7.

Averaging across the 1198 MC items in the data set, the mean item discrimination coefficient was 0.25 ± 0.14; the unweighted mean coefficient, computed by averaging together the mean coefficients for the 16 tests, was 0.27 ± 0.04. Overall, 15% of items had discrimination coefficients greater than 0.40 and thus were very strong discriminators. However, more than 30% of items were unsatisfactory discriminators, having coefficients below the benchmark value of +0.20, and 4% of items actually had negative coefficients. As Table 2 indicates, the discriminatory power of MC items varied dramatically across tests, with mean discrimination coefficients for tests ranging from a respectable 0.33 down to a rather dismal 0.20. On five tests, more than 80% of items had satisfactory discrimination coefficients, but on three tests, less than 60% of the items had satisfactory coefficients. Furthermore, on four tests, more than 6% of items had discrimination coefficients that were negative rather than positive.

Table 1
Description of Tests

	Test							
	1	2	3	4	5	6	7	8
Year Level	1	1	1	1	1	1	1	2
Faculty	AHS	AHS	AHS	SS	SS	MS	MS	BUS
No. of MC Items	60	195	211	70	30	125	36	75
No. of Examinees	266	327	269	547	458	451	371	126
Test Scores: M (SD)	52.0 (12.8)	64.7 (9.0)	66.0 (9.0)	56.7 (11.3)	58.5 (11.9)	45.2 (13.1)	63.6 (12.7)	70.6 (7.6)
Median	50.0	65.6	67.3	55.7	60.0	44.8	63.9	70.7
Range	23-87	37-87	42-83	30-96	20-90	17-85	25-92	51-87
Cronbach's alpha	0.79	0.89	0.91	0.78	0.57	0.91	0.69	0.67
Adjusted alpha[a]	0.76	0.68	0.70	0.72	0.69	0.81	0.75	0.57
	Test							
	9	10	11	12	13	14	15	16
Year Level	2	2	2	2	2	2	3	3
Faculty	SS	SS	SS	SS	SS	MS	AHS	SS
No. of MC Items	45	24	72	40	60	85	40	30
No. of Examinees	179	184	439	349	111	200	125	109
Test Scores: M (SD)	60.4 (11.7)	64.6 (14.7)	67.7 (12.2)	70.7 (11.5)	63.6 (13.4)	66.5 (14.0)	73.7 (10.5)	63.2 (10.4)
Median	60.0	66.7	68.1	70.0	63.3	67.1	75.0	63.3
Range	33-87	25-96	36-96	35-98	30-93	27-93	42-98	23-83
Cronbach's alpha	0.70	0.66	0.85	0.69	0.83	0.90	0.65	0.62
Adjusted alpha	0.73	0.80	0.79	0.73	0.80	0.84	0.69	0.73

Note: Test scores are reported as percentages.
[a]Adjusted values of Cronbach's alpha reflect a test length of 50 items.

Table 2

Item Analysis

		Test							
		1	2	3	4	5	6	7	8
No. of MC items		60	195	211	70	30	125	36	75
r_{PBIS}	M(SD)	.27 (.13)	.22 (.14)	.23 (.15)	.25 (.11)	.28 (.11)	.29 (.16)	.29 (.11)	.20 (.12)
	Range	−.11, .46	−.19, .50	−.25, .55	−.03, .49	−.03, .46	−.19, .57	0, .46	−.04, .54
	<0 (%)	3.3	8.2	7.1	1.4	6.7	6.4	0	1.3
	<0.20 (%)	26.7	40.5	42.2	35.7	16.7	25.6	22.2	49.3
p	M(SD)	.52 (.18)	.65 (.23)	.66 (.25)	.57 (.21)	.59 (.24)	.45 (.22)	.64 (.20)	.71 (.21)
	Range	.17, .86	.06, .97	.02, .99	.11, .96	.17, .92	.04, .97	.02, .94	.12, 1.00

continued on next page...

Table 2 (continued)

		Test							
		9	10	11	12	13	14	15	16
No. of MC Items		45	24	72	40	60	85	40	30
r_{PBIS}	M (SD)	.27 (.11)	.33 (.08)	.29 (.11)	.27 (.11)	.30 (.11)	.33 (.13)	.26 (.11)	.26 (.20)
	Range	.04, .44	.19, .46	.03, .49	.10, .48	−.01, .53	−.12, .55	−.04, .53	−.29, .58
	<0 (%)	0	0	0	0	1.7	2.4	2.5	3.3
	<0.20 (%)	28.9	4.2	15.3	27.5	18.3	14.1	25.0	33.0
p	M (SD)	.60 (.22)	.65 (.20)	.68 (.20)	.71 (.18)	.64 (.19)	.67 (.19)	.74 (.17)	.63 (.33)
	Range	.02, .94	.21, .95	.22, .96	.28, .98	.19, .96	.12, .98	.11, .97	0, .97

r_{PBIS}: Discrimination coefficient; *p*: Difficulty index

Cronbach's alpha, which is a measure of test reliability, also varied widely across tests, with values ranging from 0.62 to 0.91. Interpreting alpha is complicated by the fact that its magnitude is directly related to the number of test items, which varied substantially across tests. For purposes of comparison, it is therefore reasonable to adjust alpha to control for the number of test items (Bodner, 1980). As Table 1 shows, when Cronbach's alpha is adjusted to correspond to a test length of 50 items, there is much less variability in alpha across tests, with adjusted values ranging from 0.68 to 0.84. The mean discrimination coefficients for the 16 tests were strongly related to the adjusted values of alpha, r (14) = +0.88, $p < 0.001$, but not to the unadjusted values (r = +0.01).

Analysis of Distractors

As Table 3 shows, the MC items on 13 of the 16 tests had four options, and on the remaining tests, they had five. Therefore, there were 3819 distractors in the data set, and many were flawed. More than one-third (37.3%) of the distractors were flawed because they were chosen by less than 5% of examinees. In addition, 16.5% were flawed because they had a discrimination correlation that was either equal to or greater than zero (5.5% and 11.0%, respectively). In all, 45.2% of distractors had at least one of these flaws, and thus only 54.8% of distractors functioned properly.

The percentage of properly functioning distractors varied substantially across tests, ranging from a low of 37.5% (Test 15) to a high of 76.7% (Test 1). Averaging across tests, the mean number of functional distractors per item was only 1.77 ± 0.38, with means ranging from a low of 1.13 (Test 16) to a high of 2.61 (Test 6). In 11.3% of the 1198 MC items that were examined, none of the distractors functioned properly, while in 17.9% of items, all distractors functioned properly. The modal number of functional distractors was two.

We carried out one-way ANOVAs to examine the extent to which the number of functional distractors contributed to item quality. Analyses were conducted separately for the four- and five-option items, with the independent variable being the number of functional distractors contained in an item. Data for the two dependent variables, item difficulty and item discrimination, were analyzed separately. As Table 4 shows, item difficulty was strongly related to the number of functional distractors. Thus, as the number of functional distractors increased, fewer examinees selected the keyed option. This was true for both the 973 four-option items, F (3, 969) = 94.12, $p < 0.001$, $\eta^2 = .23$, and the 225 five-option items, F (4, 220) = 24.23, $p < 0.001$, $\eta^2 = .31$.

Table 3
Distractor Analysis

	Test							
	1	2	3	4	5	6	7	8
No. of MC Items	60	195	211	70	30	125	36	75
Options/Item	4	4	4	4	4	5	4	4
Total No. of Distractors	180	585	633	210	90	500	108	225
% of Distractors with								
Frequency<5%	14.4	43.4	42.7	27.6	30.0	22.4	33.3	48.4
Disc. Correlation ≥0	10.6	15.9	22.7	10.0	8.9	14.2	7.4	30.7
≥1 of the Above Flaws	23.3	51.5	52.9	34.8	38.9	34.8	37.0	59.6
Functional Distractors (%)	76.7	48.5	47.1	65.2	61.1	65.2	63.0	40.4
Functional Distractors/item: M	2.30	1.46	1.41	1.96	1.83	2.61	1.89	1.21
Functional Distractors/item: %								
None	0.0	14.4	17.1	4.3	10.0	4.0	8.3	21.3
One	16.7	40.0	35.1	22.9	16.7	11.2	16.7	40.0
Two	36.7	31.3	37.4	45.7	53.3	25.6	52.8	34.7
Three	46.7	14.4	10.4	27.1	20.0	38.4	22.2	4.0
Four	—	—	—	—	—	20.8	—	—

continued on next page...

Table 3 (continued)

	Test							
	9	10	11	12	13	14	15	16
No. of MC Items	45	24	72	40	60	85	40	30
Options/Item	4	4	4	4	5	4	5	4
Total No. of Distractors	135	72	216	120	240	255	160	90
% of Distractors with								
Frequency<5%	27.4	38.9	36.6	41.7	45.4	34.9	60.6	52.2
Disc. Correlation ≥0	16.3	9.7	5.6	6.7	19.6	11.4	30.0	27.8
≥1 of the Above Flaws	34.1	40.3	38.9	43.3	51.7	40.4	62.5	62.2
Functional Distractors (%)	65.9	59.7	61.1	56.7	48.3	59.6	37.5	37.8
Functional Distractors/Item: M	1.98	1.79	1.83	1.70	1.93	1.79	1.50	1.13
Functional Distractors/Item: %								
None	4.4	8.3	6.9	7.5	8.3	9.4	17.5	30.0
One	13.3	29.2	29.2	37.5	28.3	23.5	40.0	30.0
Two	62.2	37.5	33.3	32.5	35.0	45.9	25.0	36.7
Three	20.0	25.0	30.6	22.5	18.3	21.2	10.0	3.3
Four	—	—	—	—	10.0	—	7.5	—

Table 4 also shows that the discriminatory power of items improved dramatically as the number of functional distractors increased. This occurred in both the four-option items, F (3, 969) = 50.79, $p < 0.001$, $\eta^2 = .14$, and the five-option items, F (4, 220) = 6.40, $p < 0.001$, $\eta^2 = .10$.

Table 4
Item Difficulty and Item Discrimination as Related to the Number of Functional Distractors

Four-option items

	Number of Functional Distractors			
	0	1	2	3
n	118	297	379	179
p	.89 (.15)	.69 (.24)	.57 (.20)	.55 (.15)
r_{PBS}	.17 (.14)	.21 (.14)	.27 (.13)	.34 (.09)

Five-option items

	0	1	2	3	4
n	17	47	63	63	36
p	.80 (.30)	.72 (.19)	.54 (.21)	.44 (.18)	.42 (.12)
r_{PBS}	.19 (.16)	.28 (.13)	.26 (.14)	.29 (.13)	.37 (.09)

Values shown are M (SD). r_{PBIS}: Discrimination coefficient; *p*: Difficulty index

An interesting question concerns the relative importance of the two traits that are commonly said to render a distractor dysfunctional – that is, having a non-negative discrimination correlation and having a selection frequency less than 5%. To address this issue, we focused our attention on the data for the four-option items, which greatly outnumbered the five-option items. We first carried out an ANOVA to determine how the item discrimination coefficient was affected by the presence in the item of distractors with non-negative discrimination correlations. As Figure 2 shows, item discrimination fell dramatically as the number of distractors with non-negative discrimination correlations increased, F (3, 969) = 185.15, $p < 0.001$, $\eta^2 = 0.36$; the linear trend was statistically significant, F (1, 969) = 72.25, $p < 0.001$.

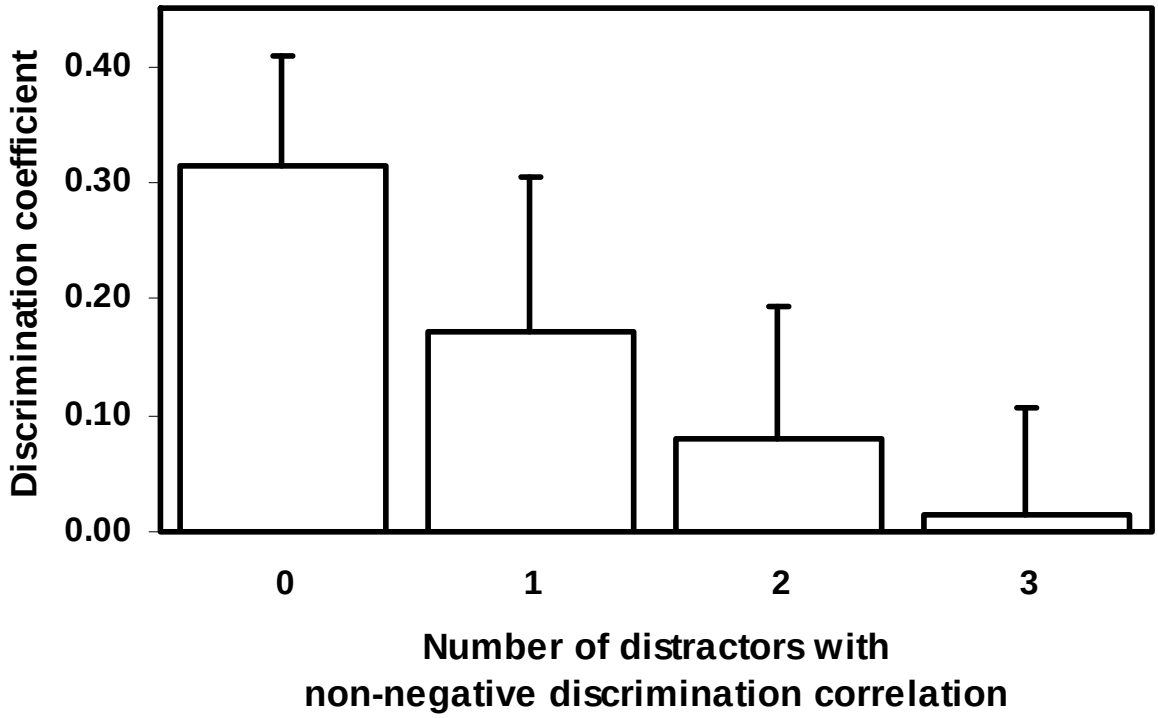

Figure 2. The effect of distractors with non-negative discrimination correlations on item discrimination. Values shown are Mean + SD. Among the four-option MC items, there were respectively 593, 305, 67, and 8 items that had 0, 1, 2 and 3 distractors with non-negative discrimination correlations.

Note that when none of the distractors had a non-negative discrimination correlation, the mean item discrimination coefficient exceeded 0.30. However, the presence of even a single distractor with a non-negative discrimination correlation had a dramatic impact on item discrimination, and when more than one such distractor was present, items lost virtually all their discriminatory power.

In contrast, the presence of distractors that were selected by less than 5% of examinees had only a minor effect on item discrimination. Figure 3 shows the mean discrimination coefficient for four-option items as a function of the number of low-frequency distractors that they contained. An ANOVA revealed a statistically significant effect, but the effect size was very small, F (3, 969) = 3.92, $p < 0.01$, $\eta^2 = 0.01$. Indeed, the presence of one or even two low-frequency distractors in an item had only a negligible effect on its discrimination coefficient. Furthermore, even those items with three low-frequency distractors still had a mean discrimination coefficient greater than 0.20. Taken together, these results show that distractors chosen by less than 5% of examinees do far less damage to item discrimination than do distractors with a non-negative discrimination correlation.

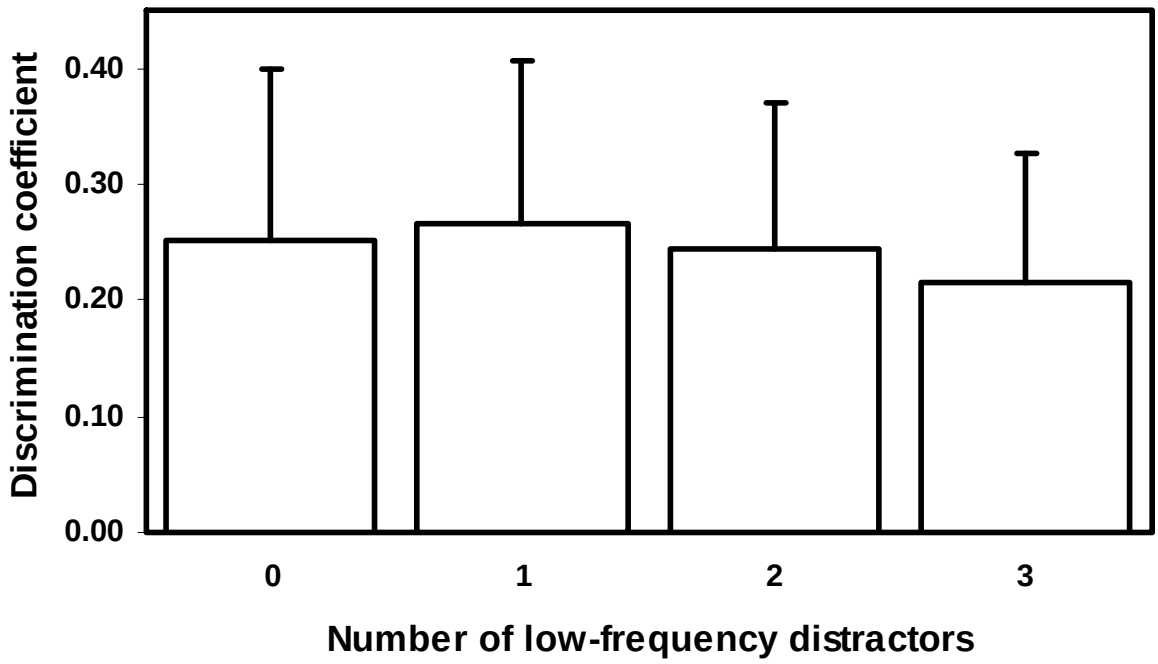

Figure 3. The effect of distractors selected by less than 5% of examinees on item discrimination. Values shown are Mean + SD. Among the four-option MC items, there were respectively 298, 340, 238 and 97 items that had 0, 1, 2 and 3 low-frequency distractors.

Relationship between Item Difficulty and Item Discrimination

Figure 4 illustrates the relationship between the difficulty index and the discrimination coefficient of all of the MC items in the data set. An ANOVA of these data indicates that the two variables are strongly related, F (9, 1188) = 31.03, $p < 0.001$, $\eta^2 = 0.19$. Examination of the figure suggests that there are both linear and quadratic components, and trend analyses confirm this, F (1, 1188) = 173.36 and 146.68 respectively, $p < 0\ .001$ in each case. The mean discrimination coefficient was very low for the most difficult items, higher for items with difficulty indices between 0.30 and 0.89, and then somewhat lower for the easiest items.

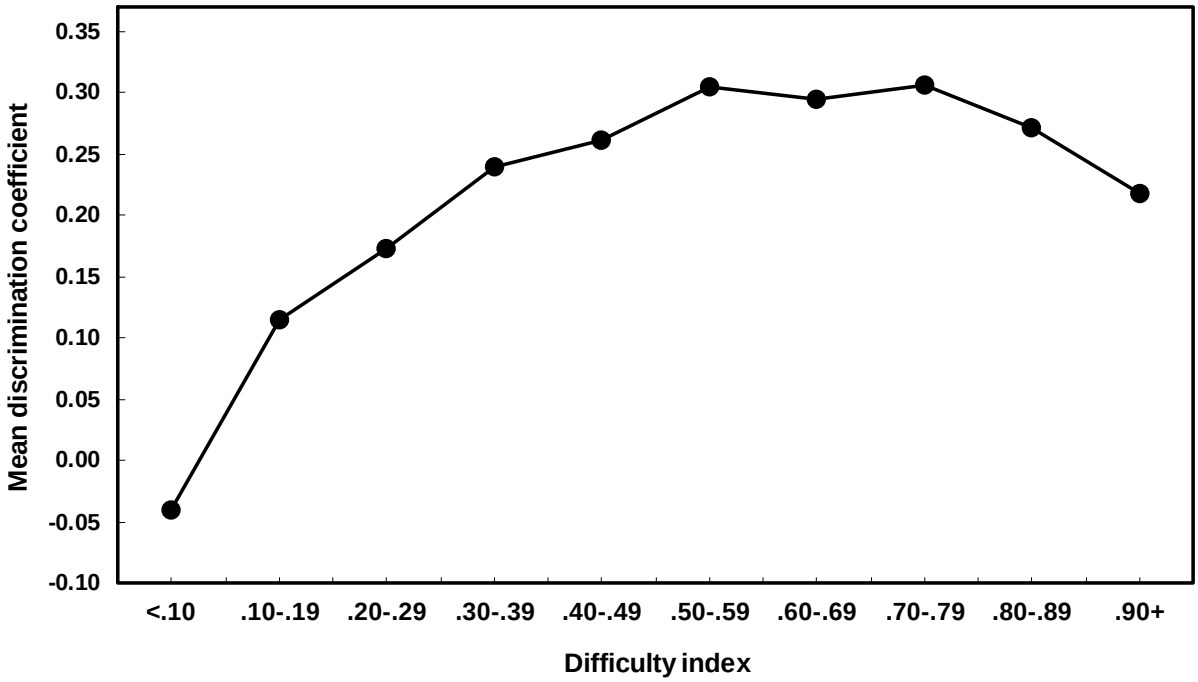

Figure 4. Mean item discrimination coefficient as a function of the item difficulty index (*p*). Standard deviations ranged from 0.10 to 0.16. Lower values represent more difficult items, and vice versa.

Figure 5 shows the percentage of items with a discrimination coefficient of 0.20 or greater as a function of item difficulty. Fewer than half of the items with a difficulty index less than 0.30 met or exceeded the commonly accepted minimal criterion for the discrimination coefficient (i.e., 0.20), and fewer than 60% of items with a difficult index greater than 0.90 did so. In contrast, more than 80% of items with a difficulty index between .50 and .80 had a discrimination coefficient of at least 0.20.

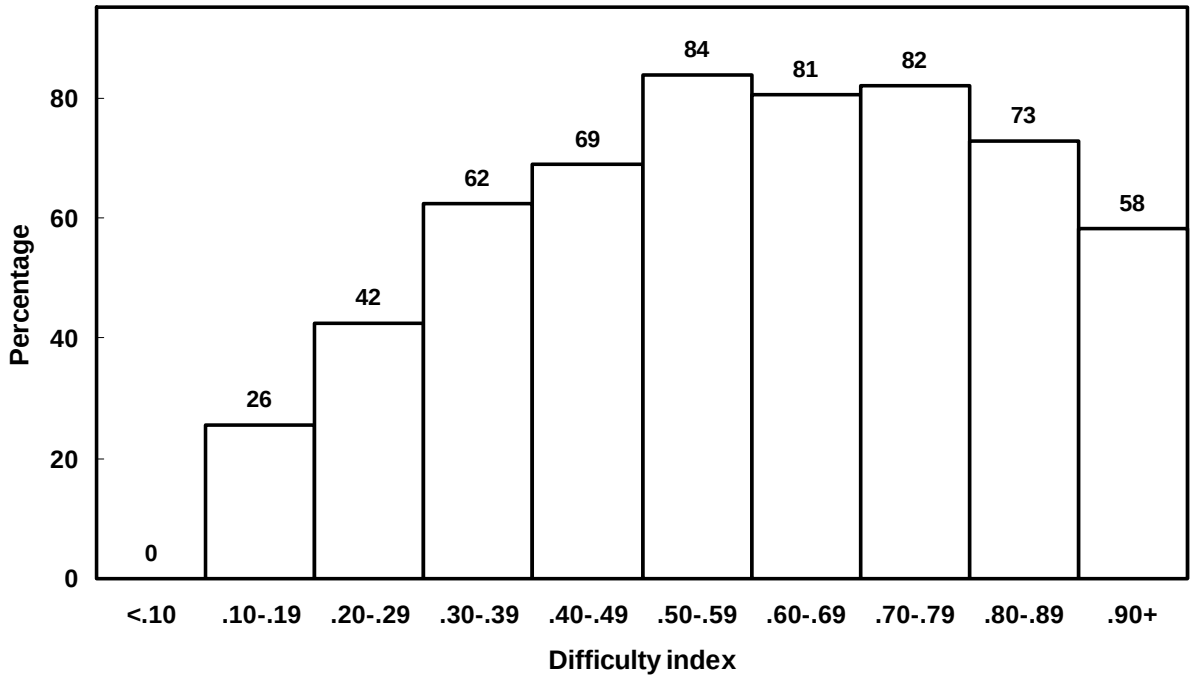

Figure 5. Percentage of items with a discrimination coefficient of +0.20 or greater as a function of the item difficulty index (*p*). Lower values represent more difficult items, and vice versa.

Discussion

The survey, which had a respectable response rate of almost 50%, revealed that MC items were used in over half of the undergraduate courses sampled. Not surprisingly, they were especially common in lower-year and in high-enrolment courses. MC items were used in about two-thirds of courses at the first- and second-year level and in the great majority of courses with enrolments of 95 or more. When instructors used MC items on their tests, they accounted for almost one-third of the total course marks. The weighting of MC items was especially heavy in larger classes, and in some courses MC items accounted for more than half of the course marks. Thus, MC items are not only widely used, but in many cases they are also a substantial determinant of students' course grades (Ross, Anderson, & Gaulton, 1987; Siegfried, Saunders, Stinar, & Zhang, 1996).

MC items were extensively used in each of the university's faculties that were studied except for Humanities, where they were used by a single instructor and counted for a modest 12.5% of course marks. Several decades ago, Cirino-Gerena (1981) pointed out that instructors in the humanities relied heavily on constructed-response techniques such as essays in their classroom testing, and the survey findings suggest that there has been little change in the use of MC testing in the humanities in the intervening years. This is not a trivial matter because evidence

suggests that in at least some disciplines within the humanities, such as music and history, constructed-response tests may be much less reliable than MC tests, while at the same time taking more time and costing far more money to grade (Wainer & Thissen, 1993). From time to time, there have been calls from within the humanities for an increased use of high-quality MC items on classroom tests (Karras, 1985; Wimmers, 1989), but these calls have apparently fallen on deaf ears and have even been actively resisted (Cohen & Rosenzweig, 2006). Given the financial challenges currently facing colleges and universities, it may be appropriate for instructors in the humanities to give serious consideration to including MC testing in their repertoire of classroom assessment techniques.

The 1198 MC items examined in this research study had a mean discrimination coefficient of 0.25. About one-sixth of the items had discrimination coefficients above 0.40 and therefore had very good discriminatory power. However, twice as many items had coefficients less than the benchmark value of +0.20. These findings, which are generally similar to those in previous studies of classroom tests (e.g., Oppenheim, 2002), suggest that there is considerable room for improvement in the MC items being used in today's classroom tests (Stiggins, 1988; Stiggins et al. 1986).

The discriminatory power of the MC items that were studied varied rather dramatically across the sixteen tests that were examined. Because the reliability of a MC test depends primarily on the discriminatory power of the test items that comprise it (Ebel, 1967), it is not surprising that the tests with lower mean discrimination coefficients also had the lowest adjusted values of Cronbach's alpha. The importance of having individual MC items with good discriminatory power and a high level of test reliability cannot be overstated. For instance, 8% of the items on Test 2 had negative discrimination coefficients. If this test were to be rescored with these badly flawed items eliminated, the scores of 96% of the examinees would increase. Moreover, the presence of flawed items on a test may particularly disadvantage better students (Downing, 2005; Tarrant and Ware, 2008). Thus, rescoring of Test 2 would lead to substantially bigger increases in test scores for examinees in the top quartile (4.36 ± 1.14 percentage points) as opposed to the bottom quartile (1.32 ± 1.22), t (162) = 16.46, $p < .001$, $d = 1.58$. With respect to reliability, Wainer and Thissen (1996) point out that when test reliability is 0.70, the test scores of 25% of students would be expected to change by more than one full standard deviation upon retesting. However, with a test reliability of 0.80, this percentage is cut in half, and at 0.90, it falls to a mere 3%. Although it is generally recommended that classroom-type assessments have a reliability of at least 0.70 (Downing, 2004), more than half of the tests we looked at had either unadjusted or adjusted values of Cronbach's alpha that fell short of this criterion. On the other hand, it must be noted that some of the tests were quite good in this regard. For example, three tests (10, 13, and 14) had mean discrimination coefficients of at least 0.30 and adjusted alpha values of 0.80 or higher.

The data revealed the expected curvilinear relationship between item difficulty and item discrimination (Ebel & Frisbie, 1991; Sim & Rasiah, 2006). Items with a difficulty index either below 0.30 or above 0.90 were less likely than other items to have satisfactory discrimination coefficients. Given the relationship between item difficulty and item discrimination, it is ironic then that some instructors like to put several very easy MC items on their tests to "make students feel good about themselves" (first author, personal observations). Although being supportive of

one's students is certainly a laudable goal, strategies that do not compromise test quality might be more appropriate for this purpose. In addition, some instructors seem to believe that very difficult MC items must be "really good" because they are so challenging and allow the best students to show off their knowledge (first author, personal observations). As it happens however, these very difficult items are often poor discriminators and may therefore detract from the overall quality of the test.

In this study, a properly functioning distractor was operationally defined as one that had a negative discrimination correlation and was selected by at least 5% of examinees. By this definition, only 55% of the distractors in this study functioned properly and the mean number of functional distractors per item was only 1.77. In addition, as the number of functional distractors in MC items increased, discriminatory power increased and items became more difficult. These findings are similar to those of Tarrant et al. (2009), who used the same operational definition for a functional distractor. Our study also revealed that the discriminatory power of items fell off sharply as the number of distractors with a non-negative discrimination correlation increased, but discriminatory power was barely affected by the presence of distractors that were selected by less than 5% of examinees. Thus, the findings indicate that the contribution that a distractor makes to an item's discriminatory power is heavily dependent on its having a negative discrimination correlation, and the frequency with which it is chosen is a much less important factor.

Assessing students' learning is an important component of the teaching process, and leaders in higher education recognize that postsecondary instructors should be able to construct classroom tests that are both reliable and valid (Smith & Simpson, 1995). Accordingly, Downing and Haladyna (1997) have emphasized the importance of improving the quality of MC items by carefully examining their discriminatory power and distractor functioning after they have been used on a test, and then modifying items so that they will function more effectively when reused in the future. A number of authors (Ebel & Frisbie, 1991; Thorndike, 2005) have recommended that MC items with a discrimination coefficient less than +0.20 be either eliminated completely or else modified before being used again, with other items also being carefully examined to see if they can be improved. In some cases, an item's unsatisfactory discriminatory power may be attributable to violation of one or more item-writing guidelines, and rewriting the item to bring it into conformity with the guidelines may be called for (Haladyna, Downing, & Rodriguez, 2002). In addition, because the quality of the distractors is so important in determining an item's discriminatory power, the modification or replacement of poorly functioning distractors will often play a key role in the improvement process. Our findings indicate that the distractors with the most detrimental effect on discriminatory power are those that have a positive discrimination coefficient, and we conclude that these distractors should be a primary focus of attention when items are being modified for future use. As mentioned earlier, guidelines for writing effective distractors are available (Haladyna, 2004; McDonald, 2007), and they may prove to be very helpful during the modification process.

This research study had several noteworthy strengths. Compared to other studies that have examined MC tests used in postsecondary classrooms, we looked at more tests, more items, and more distractors. In addition, we obtained data from across a variety of undergraduate programs, whereas previous studies have generally focused on only one discipline, and we ensured the trustworthiness of our analyses by looking only at tests that had at least 24 MC items and that

were administered to at least 100 examinees. Furthermore, we included in our analysis all available tests that met these criteria; that is, unlike Tarrant et al. (2009), we did not eliminate tests that had low reliability because this might have given a distorted view of the quality of the MC items being used on classroom tests.

A limitation of this study is that all of the data were obtained from a single mid-sized Canadian university. However, the results obtained here, which indicated that almost one-third of the MC items had unsatisfactory discriminatory power, are generally consistent with the findings of a number of previous studies (Oppenheim, 2002; Tarrant et al., 2009; Ware & Vik, 2009). It is therefore quite possible that the state of MC testing may be rather similar at other Canadian postsecondary institutions. Certainly, studies dealing with the quality of MC items on classroom tests at other postsecondary institutions, both smaller and larger, would be welcome. In the meantime, because MC items are used so often and can play such an important role in determining students' course grades, we believe that postsecondary institutions and classroom instructors should take active steps to ensure that MC testing is of high quality. Naturally, if instructors are to modify their MC items in an effort to improve their discriminatory power, they must first have access to a user-friendly item-analysis report that provides information about the discriminatory power and the distractor performance of their test items. Accordingly, we believe that postsecondary institutions have a responsibility to provide an item analysis report to instructors following every MC test that they administer. With such a report in hand, instructors will have the information that they need to allow them to work toward improving the quality of their MC tests (Su et al., 2009).

Unfortunately, most postsecondary instructors are not formally trained in the principles of testing, and only about one-third of them even understand terms such as item discrimination and reliability (McDougall, 1997). Thus, most instructors would probably not be able to interpret the information in an item analysis report, and even if they could, they would probably not know how to rewrite and improve their own MC items for future use. For this reason, we believe that postsecondary institutions must take on the responsibility of providing instructors with the training and support that they need to create high-quality MC items and to make effective use of an item analysis report for the purpose of improving items that they have used on classroom tests. Likewise, we believe that instructors have a responsibility to avail themselves of this training in order to ensure that their MC tests are well-constructed and have acceptable discriminatory power (Whitley, Perkins, Balogh, Keith-Spiegel, & Wittig, 2000). The creation of high-quality MC items is a learnable skill (Hansen, 1997; Jozefowicz et al., 2002), and by working together, institutions and instructors can improve the state of MC testing at Canadian universities.

References

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*, 31-36. http://dx.doi.org/10.1177/0273475302250570

Biggs, J. (1999). *Teaching for quality learning at university*. Buckingham, UK: Society for Research into Higher Education and Open University Press.

Bodner, G. (1980). Statistical analysis of multiple-choice exams. *Journal of Chemical Education, 57*, 188-190. http://dx.doi.org/10.1021/ed057p188

Buckles, S., & Siegfried, J. J. (2006). Using in-depth multiple-choice questions to evaluate in-depth learning of economics. *Journal of Economic Education, 37,* 48-57. http://dx.doi.org/10.3200/JECE.37.1.48-57

Cirino-Gerena, G. (1981). Strategies in answering essay tests. *Teaching of Psychology, 8*, 53-54. http://dx.doi.org/10.1207/s15328023top0801_20

Cizek, G. J., & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement, 54,* 861-872. http://dx.doi.org/10.1177/0013164494054004002

Cohen, D. J., & Rosenzweig, R. (2006). No computer left behind. *Chronicle of Higher Education, 52*(25), B6-B8.

Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian, 12,* 19-24. http://dx.doi.org/10.1016/S1322-7696(08)60478-3

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics–Physics Education Research, 5,* 020103. http://dx.doi.org/10.1103/PhysRevSTPER.5.020103

Downing, S. M. (2003). Guessing on selected-response examinations. *Medical Education, 37,* 670-671. http://dx.doi.org/10.1046/j.1365-2923.2003.01585.x

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*, 1006–1012. http://dx.doi.org/10.1111/j.1365-2929.2004.01932.x

Downing, S. M. (2005). 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. *Advances in Health Sciences Education, 10*, 133-143. http://dx.doi.org/10.1007/s10459-004-4019-5

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10,* 61-82. http://dx.doi.org/10.1207/s15324818ame1001_4

Ebel, R. L. (1967). The relation of item discrimination to test reliability. *Journal of Educational Measurement, 4,* 125-128. http://dx.doi.org/10.1111/j.1745-3984.1967.tb00579.x

Ebel, R. L. (1975). Can teachers write good true-false items? *Journal of Educational Measurement, 12,* 31-35. http://dx.doi.org/10.1111/j.1745-3984.1975.tb01006.x

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.) Englewood Cliffs, NJ: Prentice-Hall.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice item? *Educational and Psychological Measurement, 53,* 999-1010. http://dx.doi.org/10.1177/0013164493053004013

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-344. http://dx.doi.org/10.1207/S15324818AME1503_5

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business, 73*, 94-97. http://dx.doi.org/10.1080/08832329709601623

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical examinations. *Academic Medicine, 77*, 156-161. http://dx.doi.org/10.1097/00001888-200202000-00016

Karras, R. W. (1985). A realistic approach to thinking skills: Reform multiple-choice questions. *Social Science Record, 22(2),* 38-43.

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation, 4(10)*. Retrieved July 30, 2010 from http://PAREonline.net/getvn.asp?v=4&n=10.

Martinez, R. J., Moreno, R., Martin, I., & Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema, 21,* 326-330.

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40,* 25-32.

Mavis, B. E., Cole, B. L., & Hoppe, R. B. (2001). A survey of student assessment in U.S. medical schools: The balance of breadth versus fidelity. *Teaching and Learning in Medicine, 13,* 74-79. http://dx.doi.org/10.1207/S15328015TLM1302_1

McDonald, M. E. (2007). *The nurse educator's guide to assessing learning outcomes* (2nd ed.). Sudbury, MA: Jones and Bartlett.

McDougall, D. (1997). College faculty's use of objective tests: State-of-the-practice versus state-of-the-art. *Journal of Research and Development in Education, 30*, 183–93.

Oppenheim, N. (2002). Empirical analysis of an examination based on the academy of legal studies in business test bank. *Journal of Legal Studies Education, 20*, 129-158. http://dx.doi.org/10.1111/j.1744-1722.2002.tb00135.x

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? *BMC Medical Education, 7,* 49. http://dx.doi.org/10.1186/1472-6920-7-49

Phipps, S., & Brackbill, M. L. (2009). Relationship between assessment item format and item performance characteristics. *American Journal of Pharmaceutical Education, 73*(8), Article 146. http://dx.doi.org/10.5688/aj7308146

Ravenscroft, S. P., Rebele, J. E., St. Pierre, K., & Wilson, R. M. S. (2008). The importance of accounting education research. *Journal of Accounting Eduation, 26*, 180-187. http://dx.doi.org/10.1016/j.jaccedu.2009.02.002

Reid, J. C. (1970). Printed comments with item analyses. *Journal of Educational Measurement, 7,* 159-160. http://dx.doi.org/10.1111/j.1745-3984.1970.tb00710.x

Ross, A. S., Anderson, R., & Gaulton, R. (1987). Methods of teaching introductory psychology: A Canadian survey. *Canadian Psychology, 28*, 266-273. http://dx.doi.org/10.1037/h0079911

Schrecker, E. (2009). The bad old days. *Chronicle of Higher Education, 55*(40), 31.

Siegfried, J. J., Saunders, P., Stinar, E., & Zhang, H. (1996). How is introductory economics taught in America? *Economic Inquiry, 34,* 182-192. http://dx.doi.org/10.1111/j.1465-7295.1996.tb01371.x

Sim, S. M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple-choice questions of a para-clinical multidisciplinary paper. *Annals of the Academy of Medicine of Singapore, 35*, 67-71.

Smith, K. S., & Simpson, R. D. (1995). Validating teacher competencies for faculty members in higher education: A national study using the Delphi method. *Innovative Higher Education, 19,* 223-234. http://dx.doi.org/10.1007/BF01191221

Stiggins, R. J. (1988). Revitalizing classroom assessment: The highest instructional priority. *Phi Delta Kappan, 69,* 363-368.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement, 22*(4), 271-286. http://dx.doi.org/10.1111/j.1745-3984.1985.tb01064.x

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practices, 5*, 5-17. http://dx.doi.org/10.1111/j.1745-3992.1986.tb00473.x

Su, W., Osisek, P. J., Montgomery, C., & Pellar, S. (2009). Designing multiple-choice test items at higher cognitive levels. *Nurse Educator, 34*, 223-227. http://dx.doi.org/10.1097/NNE.0b013e3181b2b546

Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today, 26,* 662-671. http://dx.doi.org/10.1016/j.nedt.2006.07.006

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*, 198-206. http://dx.doi.org/10.1111/j.1365-2923.2007.02957.x

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC British Medical Education, 9*, 40. http://dx.doi.org/10.1186/1472-6920-9-40

Thorndike, R. M. (2005). *Measurement and Evaluation in Psychology and Education.* Upper Saddle River, NJ: Pearson.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118. http://dx.doi.org/10.1207/s15324818ame0602_1

Walsh, C. M., & Seldomridge, L. A. (2006). Critical thinking: Back to square two. *Nursing Education, 45,* 212-219.

Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher, 31*, 238-243. http://dx.doi.org/10.1080/01421590802155597

Whitley, B., Perkins, D., Balogh, D., Keith-Spiegel, P., & Wittig, A. (2000). Fairness in the classroom. *APS Observer, 13,* 24-27.

Wimmers, E. (1989). Questioning the text: Literary analysis and multiple-choice testing. *College Board Review, 151*, 24 -29, 39.