Journal of the Scholarship of Teaching and Learning, Vol. 16, No. 1, February 2016, pp.1-14.
doi: 10.14434/josotl.v16i1.19106

Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter

Syed Haris Ali[1], Patrick A. Carr[2], Kenneth G. Ruit[3]

Plausible distractors are important for accurate measurement of knowledge via multiple-choice questions (MCQs). This study demonstrates the impact of higher distractor functioning on validity and reliability of scores obtained on MCQs. Free-response (FR) and MCQ versions of a neurohistology practice exam were given to four cohorts of Year 1 medical students. Consistently non-functioning multiple-choice distractors (<5% selection frequency) were replaced with those developed from incorrect responses on FR version of the items, followed by administration of the revised MCQ version to subsequent two cohorts. Validity was assessed by comparing an index of expected MCQ difficulty with an index of observed MCQ difficulty, while reliability was assessed via Cronbach's alpha coefficient before and after replacement of consistently non-functioning distractors. Pre-intervention, effect size (Cohen's d) of the difference between mean expected and observed MCQ difficulty indices was noted to be 0.4 – 0.59. Post-intervention, this difference reduced to 0.15 along with an increase in Cronbach's alpha coefficient of scores obtained on MCQ version of the exam. Through this study, we showed that multiple-choice distractors developed from incorrect responses on free-response version of the items enhance the validity and reliability of scores obtained on MCQs.

Keywords: Assessment; Psychometrics; Validity; Reliability; Medical Education

Introduction

Validity of obtained scores is necessary for an assessment instrument and is irrespective of the level of examinees' education or the domain or subject under assessment. A search of the literature on scholarship of teaching and learning reveals a plethora of studies on the topic, ranging from cultural validity of assessment (Shaw, 1997), to impact of clarity of assessment's design on learners' performance (Solano-Flores & Nelson-Barber, 2001). In medical education, the desire to yield valid assessment scores is even stronger, since learner competence has immediate and serious implication on patient care. Although the study presented here was conducted in the context of undergraduate medical education, it demonstrates how the multiple-choice question, an assessment instrument prevalent in science and humanities education, can be improved to help educator scholars make more definitive conclusions about competence of learners and effectiveness of curricula.

[1] Department of Internal Medicine, University of North Dakota School of Medicine and Health Sciences, Sanford Health, 801 Broadway North, Fargo ND 58122-0114, syedharis.ali@med.und.edu
[2] Department of Basic Sciences, University of North Dakota School of Medicine & Health Sciences, 501 N Columbia Rd., Grand Forks ND 58202
[3] Department of Basic Sciences, University of North Dakota School of Medicine & Health Sciences, 501 N Columbia Rd., Grand Forks ND 58202

Validity of Scores Obtained on Multiple-choice Questions

Validity is defined as the extent to which scores obtained on an assessment instrument represent true knowledge (Cook & Beckman, 2006). To assess an exam's ability to elicit true knowledge, systematic collection of evidence of validity of assessment scores is advised (Kern et al., 2009). A source of such evidence, termed *Relations to Other Variables*, ascertains closeness of scores obtained on one instrument to scores obtained on the reference instrument for assessment of that competency. In regards to knowledge of basic medical sciences, questions written in Free-Response (FR) or un-cued (UnQ) formats have served as a point of reference for questions written in multiple-choice question (MCQ) format, since FR or UnQ formats minimize the impact of guessing and cueing (Haladyna & Downing, 1993; Damjanov et al., 1995). A synopsis of a few relevant studies follows.

Damjanov et al. did not find any significant difference between scores on or item discrimination indices between MCQ and UnQ versions of an exam and recommended un-cued open-ended format as an acceptable alternative to the MCQ format (Damjanov et al., 1995). Fajardo et al. compared performance on un-cued format of an exam with performance on its MCQ and reported a lower level of performance on un-cued version of the items (Fajardo & Chan, 1993). Prihoda et al. proposed a "correction" for random guessing for scores obtained on a MCQ exam (Prihoda et al., 2006). The correction entailed a weighting formula for points awarded for correct answers, incorrect answers, and unanswered questions such that the expected value of increase in test score due to guessing was zero. They reported that the agreement between scores obtained on FR version of the exam and "corrected" MCQ scores was greater than the "uncorrected" scores, highlighting the value of correction for guessing in validity of scores obtained on MCQs. Newble et al. compared performance of medical students and practicing physicians on a test of clinical knowledge written in MCQ and FR formats (Newble et al., 1979). They reported a smaller difference between mean scores obtained on the two versions among practicing physicians than among senior-level and junior-level students, surmising that MCQs performance appears to overestimate examinee ability which makes them less suitable for assessment of clinical competence.

The difference between performance on an item's FR and MCQ format can be attributed to functioning distractors (Haladyna & Downing, 1993; Rodriquez, 2005). By definition, a *functioning* distractor (FD) is an incorrect option selected by ≥5% of examinees (i.e., ≥5% selection frequency), and chosen by a greater number of low-performing examinees than high-performing ones, which renders a negative discriminatory ability to that distractor (Rodriguez, 2005). On the other hand, a *non-functioning* distractor (NFD) does not possess these desirable characteristics. Tarrant et al. reported on the impact of eliminating a non-functioning distractor from 4- or 5-option MCQs and reported minimal decrease (0.3%) in mean item difficulty (Tarrant & Ware, 2010). They reported that three-option version of the items contained more functioning distractors despite having fewer distractors overall, and that existing distractors more discriminatory after removal of infrequently selected distractors. A seminal study published by Rodriguez consolidated the findings from dozens of previously published studies and showed that systematically removing one non-functioning distractor from 5-option MCQs reduced their average difficulty and discriminatory ability only to a mild extent (0.02 and 0.04 units, respectively), and removing two non-functioning distractors from such questions did not impact average item discriminatory ability (Rodriguez, 2005). The above studies show that non-functioning distractors offer very little in terms of validity of scores, while unnecessarily increasing the response time needed per MCQ.

Reliability of Scores Obtained on Multiple-choice Questions

The concept of reliability is ingrained in Classical Test Theory, the central tenet of which is that an examinee's observed score (X) can be decomposed into her/his true score (T) and a random error component (E) (X = T + E) (De Champlain, 2010). True score (T) is the score obtained if the exam were measuring the ability of interest perfectly (i.e. with no measurement error). A reliability coefficient, which ranges from 0 (lowest) to 1 (highest), estimates of the level of concordance between observed and true scores of an examinee (De Champlain, 2010).

The type of reliability frequently discussed in the context of MCQ is *internal consistency*, which is meant for exams that require a single administration to a group of examinees (Downing, 2004). Internal consistency reliability assesses the correlation between scores obtained on two parallel forms of an exam, i.e., the forms assessing the same content and on which examinees have the same true scores and equal errors of measurement. Cronbach's alpha is its widely-used coefficient; a coefficient of 0.8 or more is desired for high-stakes in-house exams (De Champlain, 2010; Downing, 2004).

It has been suggested that reliability can be improved by increasing the number of items given in an exam (Downing, 2004). Such an improvement can be estimated using the Spearman-Brown "prophecy" formula $\alpha = \frac{k}{k-1}\,(1 - \frac{sum\ of\ variances\ of\ all\ items}{total\ test\ variance})$, where "$\alpha$" is the Cronbach's alpha coefficient and "k" is the number of items in an exam (Karras, 1997). However, owing to the usually fixed number of items given in high-stakes in-house or licensure exams, an alternate way to improve reliability is to increase the spread of scores obtained on an exam (total test variance). An increased distribution of scores can be obtained by eliciting a wider range of performances from examinees by giving a greater number of moderately difficult (difficulty index: 0.4 – 0.8) and sufficiently discriminatory (point biserial correlation ≥ 0.2) items in the exam (Hutchinson et al., 2002). McManus et al. discuss in greater detail how this approach may increase the standard deviation, hence variance, of observed scores (McManus et al., 2003).

In the study presented here, two versions (FR and MCQ) of the same neurohistology exam were randomly distributed among six cohorts of Year 1 medical students. The evidence of validity pertaining to *Relations to other variables*, described above, was gathered before and after replacement of consistently non-functioning distractors with those developed from incorrect responses on the FR version of the items. Specifically, an index of *expected* MCQ difficulty was calculated (see Methods) and compared with the index of *observed* MCQ difficulty. This comparison was based on assumptions that, 1. FR version of an item elicits true knowledge, and 2. Faculty responsible for the assessment of basic science content writes reasonably plausible MCQ distractors. The effect of distractor functioning on range of ability elicited from examinees and its impact on reliability of obtained scores was also studied.

Research hypotheses

Research hypothesis of the *validity* part of the study was: There is no difference between *expected* and *observed* MCQ difficulty indices when selection of all provided options is accounted for in calculating the *expected* index. To date, no such comparisons of actual performance on multiple-choice questions (*observed* difficulty index) with what it ought to have been (*expected* difficulty index) have been reported, especially in the context of assessment in undergraduate medical education, which highlights the novelty of the presented study. Research hypothesis of the *reliability* part of the study was: Enhanced distractor functioning increases the standard deviation and, therefore, reliability coefficient of scores obtained on multiple-choice exams.

Materials and Methods

Research Design

An experimental research design with random distribution of the free-response (FR) and multiple-choice (MCQ) versions of an exam was employed. The study was approved and adjudged exempt from detailed review by the Institutional Review Board of University of North Dakota.

Subjects and Setting

Six cohorts of Year 1 medical students at the University of North Dakota School of Medicine and Health Sciences served as subjects.

The school's medical education curriculum is a hybrid of Patient-Centered Learning (PCL) as well as traditional, discipline-based instruction. Neurohistology is taught during the neuroscience curricular block scheduled at the end of academic Year 1 via a combination of lectures and laboratory exercises by faculty with expertise in neuroscience.

Sample of Questions

A neurohistology exam comprising 25 items with a mix of knowledge (factual recall) and application-type questions was used. A FR (fill-in-the-blank) and a MCQ (one-best answer) version of this exam was created; the only difference between these two versions was in the format of the asked question (example: Figure 1). Of the 25 FR-MCQ item-sets, two were excluded from analysis since their FR version contained options, thereby not meeting the criterion needed for comparison with the MCQ version.

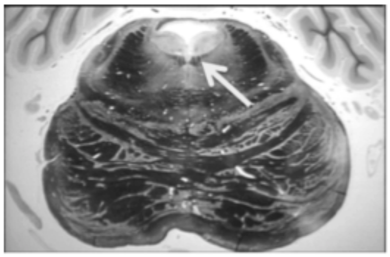
Free-response (FR) version of sample item: Name the clinical finding that would be associated with an infarct in the indicated structure.		Multiple-choice (MCQ) version of sample item: A lesion of the indicated structure will result in which of the following clinical findings? A. Lateral strabismus B. Bilateral horizontal nystagmus C. Internuclear ophthalmoplegia D. Dilated, unreactive pupil E. Central scotoma

Figure 1. Example of Free-Response (FR) and Multiple-Choice (MCQ) version of an item.

Procedure

Each cohort of students was invited, via email, to attend a non-mandatory practice session 5 days prior to the end-of-block neurohistology exam. No information in regards to design of the study was shared in advance. No points were granted for participation in the study. Once seated, an approximately equal number of free-response and multiple-choice versions of the exam printouts were randomly distributed amongst the subjects. Then, the purpose of the study was shared, and subjects were asked not to provide any personal or identifiable information on the answer sheets. Neurohistology images (example: Figure 1) were projected on a screen and one minute was provided to answer each question. After the exam, each question was discussed openly and students were asked not to change their answers. The answer sheets were collected, codified and scored according to pre-developed answer keys.

Intervention

The following revisions were performed on the MCQ version of the exam based on examinee performance in Cohorts 1 – 4.

a. Thirty-one distractors in 15 MCQs with consistent selection frequency of 0% were replaced with new distractors developed from frequent incorrect responses on FR version of the items.
b. Five 5-option MCQs were converted to 4-option MCQs via removal of a distractor with consistently 0% selection frequency. The number of 5-, 4- and 3-option MCQs in the original (unrevised) version was 21, 1 and 1, respectively; these numbers were 16, 6 and 1 in the revised MCQ version of the exam.

In order to note the extent of distractor functioning from a bigger sample of subjects, the revised MCQ version of the exam was given to all subjects in Cohort 5. In Cohort 6, the revised MCQ version was given to random half of subjects while the other half received the FR version of the exam.

Data Collection and Analysis

The following variables were calculated from student performance:

a. Individual, as well as mean and standard deviation of scores in each cohort.
b. Psychometric characteristics, i.e. the difficulty and discriminatory ability of each item. Difficulty was calculated via difficulty index (number of correct answers / number of all answers), while discriminatory ability was calculated via point biserial (item-total) correlation (Tavakol & Dennick, 2011).
b. The index of *expected* MCQ difficulty was calculated as follows. Suppose the FR version of an item is correctly answered by 60% examinees (FR difficulty index: 0.6). The proportion of examinees with an incorrect answer on the FR version would be 40% (0.4). Now suppose that the MCQ version of this item contains 5 options. It will be anticipated that a certain proportion of examinees who answered the item incorrectly on its FR version might have chosen the correct MCQ option, using random or educated guessing, had they taken the MCQ version of the exam. Probability would suggest that such a proportion among 40% (0.4) examinees would be at least 8% (0.08) (0.4 / 5 = 0.08). This proportion of examinees (0.08) can be added to the FR difficulty index to generate the index of *expected* MCQ difficulty (0.6 + 0.08 = 0.68) (Table 1).

Table 1. Calculation of *expected* MCQ difficulty index

MCQ ID	# of total options	FR version difficulty (FR diff.)	Proportion of students with incorrect answers on FR version (Pw)	Expected inflation in item ease (EI) (Pw / # of total options in the MCQ version)	*Expected* MCQ difficulty (FR diff. + EI)
Example	5	0.60	1 – 0.60 = 0.40	0.40 / 5 = 0.08	0.60 + 0.08 = 0.68

c. Effect size [Cohen's *d*] of the difference between mean *expected* and *observed* MCQ difficulty indices. Effect size represents the extent to which research hypothesis is considered to be true, or the degree to which findings of an experiment have practical significance in the study population regardless of the size of the study sample (Hojat & Xu, 2004). Cohen's *d* is a statistic that is equal to the difference between means of experimental (M_e) and control (M_c) groups divided by the standard deviation for the control group (σ_c) (Cohen's $d = \frac{Me - Mc}{\sigma c}$) (Hojat & Xu, 2004).

e. Number of MCQ distractors with ≥5%, ≥10%, ≥20%, and ≥33% selection frequency in each cohort.

f. Cronbach's alpha coefficient of scores, before and after revision, on MCQ version of the exam.

g. Standard Error of Measurement (SEM = SD$\sqrt{1 - reliability}$), which is the standard deviation of an examinee's observed score, given her true score (Karras, 1997). SEM describes precision of measurement and is used to establish a confidence interval within which an examinee's true score is expected to fall[4].

Exam performance data from all cohorts were stored in Microsoft Excel (2010) and analyzed via MS-Excel and SigmaStat v. 20.

Results

Table 2 displays the number of students taking the FR and MCQ versions of the exam, score means and their standard deviations, mean item difficulty indices and mean point biserial correlations. As expected, scores on FR version tended to be lower in all cohorts than scores on MCQ version of the exam. Moreover, the revised MCQ version (Cohorts 5 and 6) exhibited greater difficulty and discriminatory ability than the original MCQ version (Cohorts 1 – 4) of the exam.

Table 2. Number of students taking the Free-Response (FR) and Multiple-Choice (MCQ) versions of the exam in all cohorts. Mean score, standard deviation, mean item difficulty (diff.) and mean point biserial correlations (pbi) are also displayed.

	Cohort 1		Cohort 2		Cohort 3		Cohort 4		Cohort 5	Cohort 6	
	FR	MCQ	FR	MCQ	FR	MCQ	FR	MCQ	MCQ (only)	FR	MCQ
# of students	28	31	27	31	30	23	28	27	71	34	33

[4] Standard Error of Measurement is not to be confused with another commonly used statistic Standard Error of the Mean (a.k.a. Standard Error), which is standard deviation of the sample mean's estimate of a population mean (Harvill, 1991).

Mean score	16.10	19.51	15.51	19.00	14.70	18.80	15.90	19.60	17.04	15.65	18.24
Standard deviation	3.15	3.34	4.16	2.52	3.69	2.11	4.34	2.48	3.61	3.37	3.61
Mean item diff.	0.70	0.85	0.67	0.83	0.64	0.82	0.69	0.85	0.74	0.68	0.79
Mean pbi	0.30	0.43	0.43	0.29	0.35	0.25	0.42	0.30	0.38	0.34	0.39

Table 3 and Figure 2 display Effect Size (Cohen's *d*) of the difference between mean *expected* and *observed* MCQ difficulty indices before (Cohorts 1 – 4) and after (Cohort 6) replacement of previously non-functioning distractors; Cohen's *d* could not be calculated for Cohort 5, since all subjects in that cohort received the revised MCQ version of the exam. Considerable increase in MCQ difficulty was noted after replacement of consistently non-functioning distractors (Cohorts 5 and 6), with a concomitant reduction in disparity between mean *expected* and *observed* MCQ difficulty indices (Cohort 6) (*d* = 0.15).

Table 3. Mean Free-Response (FR) and Multiple-Choice (MCQ) difficulty indices and their Standard Deviations (SD) in all cohorts. Effect size (Cohen's *d*) of the difference b/w Mean *Observed* and *Expected* MCQ difficulty indices is also displayed.

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Cohort 6
Mean FR diff. index (SD)	0.7 (0.20)	0.67 (0.25)	0.64 (0.23)	0.69 (0.18)	-	0.68 (0.22)
Mean *Expected* MCQ diff. index (SD)	0.78 (0.15)	0.76 (0.19)	0.74 (0.17)	0.77 (0.14)	-	0.76 (0.16)
Mean *Observed* MCQ diff. index (SD)	0.85 (0.13)	0.83 (0.17)	0.82 (0.19)	0.85 (0.14)	0.74 (0.16)	0.79 (0.16)
Cohen's *d* of the difference b/w Mean *Observed* and *Expected* MCQ difficulty indices	0.46	0.40	0.46	0.59	-	0.15

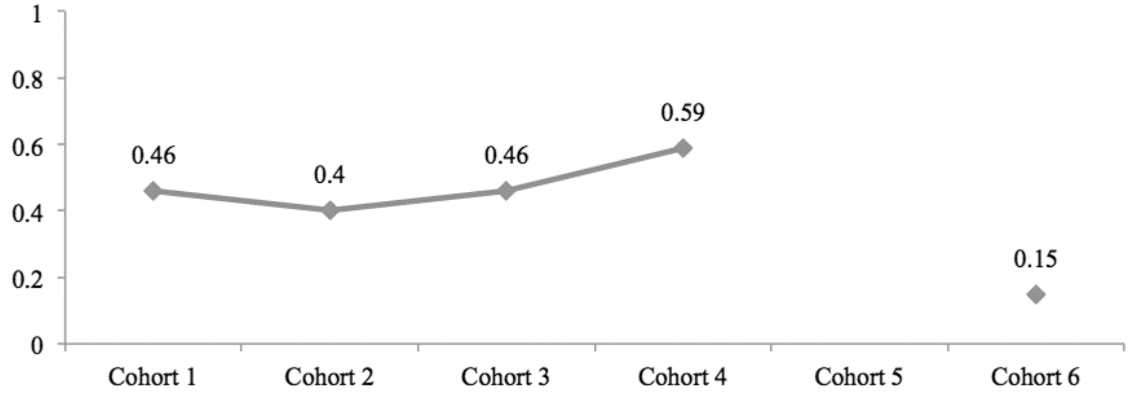


Figure 2. Effect size of the difference between mean expected and observed MCQ difficulty indices.

Table 4 and Figure 3 display the number of distractors with ≥5%, ≥10%, ≥20%, and ≥33% selection frequency in MCQ version of the exam before (Cohorts 1 – 4) and after (Cohorts 5 and 6) replacement of consistently non-functioning distractors. Table 4 also displays the number of total as well as functioning (≥5% selection frequency) distractors per MCQ. Both higher distractor selection in most categories and a greater number of functioning distractors per MCQ was noted after replacement of consistently non-functioning distractors (Cohorts 5 and 6).

Table 4. Number of distractors with ≥5%, ≥10%, ≥20% and ≥33% selection frequency in each cohort. Number of total and functioning (≥5% sel. freq.) distractors per MCQ is also displayed.

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Cohort 6
# of total distractors	89	89	89	89	84	84
≥5% sel. freq.	24 (26.97%)	22 (24.72%)	21 (23.60%)	20 (22.47%)	38 (45.24%)	23 (27.38%)
≥10% sel. freq.	6 (6.74%)	13 (14.61%)	11 (12.36%)	12 (13.48%)	21 (25.00%)	14 (16.67%)
≥20% sel. freq.	2 (2.25%)	5 (5.62%)	6 (6.74%)	4 (4.49%)	5 (5.95%)	7 (8.33%)
≥33% sel. freq.	0.00%	1 (1.12%)	5 (5.62%)	0.00%	0.00%	3 (3.57%)
# of distractors per MCQ	3.87	3.87	3.87	3.87	3.84	3.84
# of *functioning* distractors per MCQ	1.04	0.96	0.91	0.87	1.65	1.00

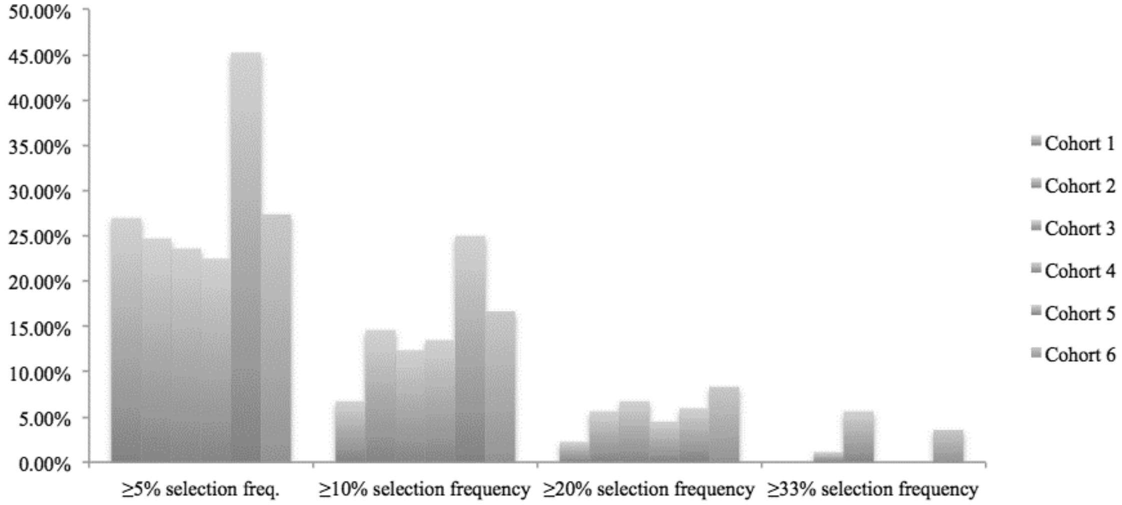


Figure 3. Percentage of MCQ distractors with different selection frequencies.

Table 5 and Figure 4 display the reliability coefficients (Cronbach's alpha) and Standard Errors of Measurement (SEM) of scores obtained on FR and MCQ versions of the exam. After replacement of previously non-functioning distractors (Cohorts 5 and 6), scores obtained on the MCQ version of the exam exhibited greater standard deviation (3.61), higher Cronbach's alpha coefficient (0.74 and 0.78) and a slightly higher Standard Error of Measurement (1.84 and 1.66). Figure 4 demonstrates the directly proportional relationship between standard deviation and reliability coefficient of exam scores. A peculiar finding was high standard deviation and reliability coefficient of scores on MCQ version of the exam in Cohort 1. This is an interesting finding, since examinees in that cohort had received the unrevised MCQ version of the exam. See Discussion for a possible explanation of this finding.

Table 5. Mean score, Standard Deviation (SD), reliability coefficient (Cronbach's alpha) and Standard Error of Measurement (SEM) on FR and MCQ versions of the exam in all cohorts.

	Cohort 1		Cohort 2		Cohort 3		Cohort 4		Cohort 5	Cohort 6	
	FR	MCQ	FR	MCQ	FR	MCQ	FR	MCQ	MCQ (only)	FR	MCQ
# of	28	31	27	31	30	23	28	27	71	34	33

students										
Mean score	16.10	19.51	15.51	19.00	14.70	18.80	15.90	19.60	17.04	15.65	18.24
SD	3.15	3.34	4.16	2.52	3.69	2.11	4.34	2.48	3.61	3.37	3.61
Range	9 – 23	8 – 23	4 – 21	12 – 23	5 – 21	14 – 23	4 – 22	13 – 23	7 – 23	7 – 22	9 – 23
Cronbach's alpha	0.63	0.80	0.82	0.61	0.73	0.43	0.81	0.64	0.74	0.67	0.78
SEM	1.91	1.49	1.75	1.56	1.91	1.59	1.88	1.49	1.84	1.87	1.66

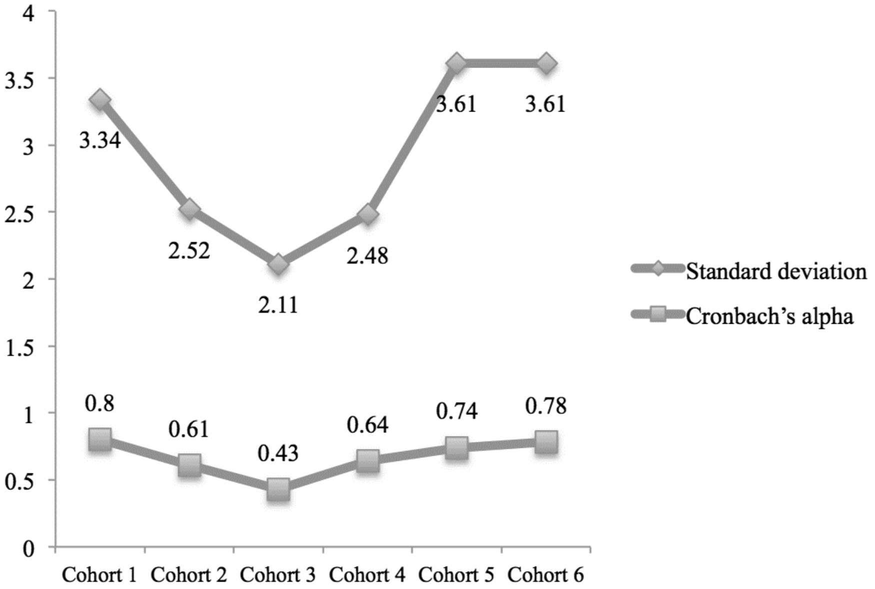


Figure 4. Standard deviation and reliability coefficient (Cronbach's alpha) of scores obtained on MCQ version of the exam.

Discussion

The first observation, in line with previously published studies (Ward, 1982; Norman et al., 1987; Schuwirth, 1996; Norman, 1988), was that performance on FR version of an exam is consistently lower than performance on its MCQ version (Table 1). Since FR and MCQ versions were randomly distributed in each cohort, the consistently disparate performance is attributable to the nature of the two versions; the MCQ version contains options and allows for some degree of cueing and correct guessing, while the FR version requires production of an answer spontaneously from memory.

Secondly, the difficulty of a MCQ-based exam is lower than expected when the number of distractors with sufficient plausibility (≥5%, ≥10%, ≥20% and ≥33% selection frequencies) is low. Tables 3 and 4 highlight this finding. Effect size of the difference between mean *expected* and *observed* MCQ difficulty indices was found to be higher in cohorts with lower overall distractor functioning (Cohorts 1 – 4). However, when consistently non-functioning distractors were replaced with those developed from frequent incorrect answers on FR version of the items (Cohorts 5 and 6), a higher overall distractor functioning and reduced disparity between mean *expected* and *observed* MCQ difficulty indices was noted. In other words, when incorrect responses on FR versions of the items are used to construct MCQ distractors, the MCQs tend to demonstrate their expected difficulty thereby enhancing the evidence of validity of scores obtained on them. This argument is strengthened by previously published reports that MCQ difficulty is contingent upon quality, not quantity, of its distractors (Tarrant & Ware, 2010; Schuwirth et al., 1996). Therefore, we surmise that careful creation and selection of distractors is vital for reducing the much-dreaded cueing effect and amelioration of quality of MCQ testing.

Unlike a previously-published study which demonstrated enhanced validity of scores obtained on MCQs via post-hoc correction for guessing (Prihoda, 2006), the study presented here used an active intervention in the form of replacement of consistently non-functioning distractors with more plausible, functioning ones. We used incorrect responses on FR version of the items in their previous no-stakes administration, which is an approach yet to be reported in educational research in medicine and other sciences, which highlights the novelty of the presented study.

Thirdly, replacement of consistently non-functioning distractors led to an increase in average discriminatory ability of MCQs. Table 2 highlights this finding; average point biserial correlations were found to range between 0.25 – 0.40 before revision of the MCQ version of the exam (Cohorts 1 – 4). After the revision, point biserial correlations of 0.38 (Cohort 5) and 0.39 (Cohort 6) were noted, which are considerably higher than those in the previous three cohorts. This increase in discriminatory ability occurred in the setting of increased selection, i.e. functioning, of MCQ distractors (Table 4, Figure 3) and affirms the notion that plausible distractors gauge conceptual misunderstandings more accurately, allowing clearer separation of low- and high-ability examinees.

Fourthly, increased distractor functioning enhances the reliability coefficient of scores obtained on MCQs (Table 5). After replacement of consistently non-functioning distractors (Cohorts 5 and 6), performance on the MCQ version of the exam exhibited a lower mean, greater range, and higher standard deviation of scores. Owing to the directly proportional relationship between standard deviation and the reliability coefficient (Karras, 1997), the higher standard deviation led to an increase in the reliability of scores as well. This finding highlights the effect of enhanced distractor functioning on spread (standard deviation) of scores and, consequently, on the reliability coefficient of scores obtained on an exam.

Looking at the data presented in Table 5, a noteworthy finding is the relatively higher standard deviation (3.34) and reliability coefficient (0.84) on MCQ version of the exam in Cohort 1. These values are higher than other cohorts that also took the original (unrevised) MCQ version of the exam (Cohorts 2 – 4). Scores obtained on the MCQ version of the exam in Cohorts 1 – 4 were weakly valid as evidenced from a relatively higher effect size of the difference between *expected* and *observed* MCQ difficulty indices (Table 2). Therefore, a higher reliability coefficient in Cohort 1 was puzzling to us. However, a closer look at range of scores explains this finding. While maximum score on the MCQ version of the exam was the same across Cohorts 1 – 4, a minimum score of 8 was observed in Cohort 1, while it was 12, 14 and 13 in Cohorts 2, 3 and 4, respectively. This shows that, for some reason, there were some very low performing examinees in Cohort 1 who took the MCQ version of the exam. Their uncharacteristically lower performance increased the score range, the standard deviation, and, consequently, the reliability coefficient of scores on MCQ version of the exam, owing to the directly proportional relationship between standard deviation and the reliability coefficient (Karras, 1997).

Another peculiar finding was a slight increase in the standard error of measurement (SEM) on the MCQ version of the exam after replacement of consistently non-functioning distractors (Cohorts 5 and 6) (Table 5). The explanation for this finding is the directly proportional relationship between standard deviation (SD) and standard error of measurement

(SEM) (SEM = SD$\sqrt{1-reliability}$) (Hutchinson et al., 2002; Harvill, 1991), since increased range of ability (standard deviation) elicited by an exam increases not only the reliability coefficient but also the error of measurement of assessment instrument. This theory has been reported on by Tighe et al., who studied the interrelationships among standard deviation, Standard Error of Measurement and exam reliability via a Monte Carlo simulation of 10,000 candidates taking a postgraduate exam (Tighe et al., 2010). They found that scores obtained on the very same exam experienced a decrease in reliability coefficient when retaken by only those examinees who had already passed it. In other words, allowing very weak (unprepared) candidates to take an exam can artificially inflate the reliability of scores obtained on an exam. Tighe et al. suggested that, when ability range of examinees is noted to be narrow, the Standard Error of Measurement may be enough for assessment of measurement precision. We agree with this suggestion and advise interpretation of the reliability coefficient in light of the psychometric characteristics (difficulty index and point biserial correlations) distractor functioning of MCQs.

A number of limitations apply to the presented study. First is the small number of investigated items (n=23). Although suitable for assessment of knowledge of neurohistology, this number may be insufficient for an experiment of this nature and study may be expanded to include more items. Second potential limitation is the no-stakes nature of the exam used in this study; it was given as practice for the high stakes neurohistology exam. Despite the no-stakes nature of the experiment, it is worth noting that our research question focused solely on differential performance on FR and MCQ version of an exam at single time-points. Thirdly, a potential limitation is the generalizability of our findings. Although we anticipate considerable generalizability of our findings owing to the nature of our intervention (using common responses on FR version of the items as distractors on MCQ version of the same items), we are yet to see a replication of our experimental design in settings other than undergraduate medication education. We invite educator scholars in sciences and humanities to replicate our design and study the validity and reliability of scores obtained on MCQs revised on the principle elicited in this study. We predict that many educator scholars will find this approach to be resource-friendly and efficient.

For its ease of administration and objective grading, multiple-choice testing is the prevalent form of assessment in science and humanities education. However, it relies on recognition of the most credible answer from a brief list of options, some of which may be barely plausible. The examination is a far cry from real-life situations healthcare, science and humanities professionals face every day. Novel problems in any discipline are rarely solved simply by choosing from among a limited list of presented options. For example, in a healthcare setting, although signs and symptoms of an illness allow for some cueing and educated guessing, patients do not present the healthcare provider with five options from among which the "single best answer" is chosen (Veloski et al., 1999). In that setting, the "single best answer" is expected to be chosen based on knowledge, analysis and reason. Therefore, it is imperative that multiple-choice questions undergo strict scrutiny for their ability to elicit true knowledge. Therefore, it is imperative that multiple-choice questions undergo strict scrutiny for their ability to elicit true knowledge. Using an adequate yardstick for comparison, such as performance on open-ended, free-response version of the same questions, is a useful step in this direction and helps assess the validity of scores obtained on such questions. In medicine, licensure bodies such as National Board of Medical Examiners recognize the importance of conducting such comparisons, and a few studies of this nature have been published in the past (Case & Swanson, 1994; Swanson et al., 2005; Swanson et al., 2006). In our experience, administering two versions (free-response and multiple-choice) of the same exam as practice for a high-stakes multiple-choice exam allows learners to detect areas of needed improvement, and instructors to encourage deep, rather than superficial, learning strategies. An attempt to improve the ability of MCQs to accurately serve their purpose, through such ventures, may truly be worthy of faculty time and effort.

Acknowledgements

The authors would like to thank Linda Olson, EdD, Associate Professor, Department of Education Resources, University of North Dakota School of Medicine & Health Sciences for her support and advice throughout the project. They also extend their gratitude to Danielle Condry, PhD, Assistant Professor, Department of Basic Sciences, University of North Dakota School of Medicine & Health Sciences, and Elizabeth Sandquist, PhD, HHMI Postdoctoral Fellow at Iowa State University, for their input on the research manuscript.

References

Case, S. M., Swanson, D. B., & Ripkey, D. R. (1994). Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. Academic Medicine: *Journal of the Association of American Medical Colleges, 69* (10 Suppl), S1-3.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119* (2), 166.e7-166.16. doi:S0002-9343(05)01037-5 [pii]

Damjanov, I., Fenderson, B. A., Veloski, J. J., & Rubin, E. (1995). Testing of medical students with open-ended, uncued questions. *Human Pathology, 26* (4), 362-365.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44* (1), 109-117. doi:10.1111/j.1365-2923.2009.03425.x [doi]

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38* (9), 1006-1012. doi:10.1111/j.1365-2929.2004.01932.x [doi]

Fajardo, L. L., & Chan, K. M. (1993). Evaluation of medical students in radiology. written testing using uncued multiple-choice questions. *Investigative Radiology, 28* (10), 964-968.

Haladyna, T.M., & Downing, S.M. (1993). How many options is enough for a multiple-choice test item? *Educational Measurement: Issues and Practice. 53*, 999–1009.

Harvill, L.M. (1991). NCME Instructional module: standard error of measurement. *Educational Measurement: Issues and Practice. 10* (2), 33–41.

Hojat, M., & Xu, G. (2004). A visitor's guide to effect sizes: Statistical significance versus practical (clinical) importance of research findings. *Advances in Health Sciences Education: Theory and Practice, 9* (3), 241-249. doi:10.1023/B:AHSE.0000038173.00909.f6 [doi]

Hutchinson, L., Aitken, P., & Hayes, T. (2002). Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Medical Education, 36* (1), 73-91. doi:1120 [pii]

Karras, D. J. (1997). Statistical methodology: II. reliability and variability assessment in study design, part A. *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine, 4* (1), 64-71.

Kern, D.E., Thomas, P.A., & Hughes, M.T. (2009). Curriculum Development For Medical Education: a Six Step Approach, second edition. Baltimore: The Johns Hopkins University Press.

McManus, I. C., Mooney-Somers, J., Dacre, J. E., Vale, J. A., MRCP(UK) Part I Examining Board, & Federation of Royal Colleges of Physicians, MRCP(UK) Central Office. (2003). Reliability of the MRCP(UK) part I examination, 1984-2001. *Medical Education, 37* (7), 609-611. doi:1568 [pii]

Newble, D. I., Baxter, A., & Elmslie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education, 13* (4), 263-268.

Norman, G. R. (1988). Problem-solving skills, solving problems and problem-based learning. *Medical Education, 22* (4), 279-286.

Norman, G. R., Smith, E. K., Powles, A. C., Rooney, P. J., Henry, N. L., & Dodd, P. E. (1987). Factors underlying performance on written tests of knowledge. *Medical Education, 21* (4), 297-304.

Prihoda, T. J., Pinckard, R. N., McMahan, C. A., & Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education, 70* (4), 378-386. doi:70/4/378 [pii]

Schuwirth, L. W., van der Vleuten, C. P., & Donkers, H. H. (1996). A closer look at cueing effects in multiple-choice questions. *Medical Education, 30* (1), 44-49.

Shaw, J.M. (1997). Threats to the validity of science performance assessments for English language learners. *Journal of Research in Science Teaching, 34*, 721–743.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38*, 553–573.

Swanson, D. B., Holtzman, K. Z., Allbee, K., & Clauser, B. E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine : Journal of the Association of*

American Medical Colleges, 81 (10 Suppl), S52-5. doi:10.1097/01.ACM.0000236518.87708.9d [doi]

Swanson, D. B., Holtzman, K. Z., Clauser, B. E., & Sawhill, A. J. (2005). Psychometric characteristics and response times for one-best-answer questions in relation to number and source of options. *Academic Medicine : Journal of the Association of American Medical Colleges, 80* (10 Suppl), S93-6. doi:80/10_suppl/S93 [pii]

Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurse Education Today, 30* (6), 539-543. doi:10.1016/j.nedt.2009.11.002 [doi]

Tavakol, M., & Dennick, R. (2011). Post-examination analysis of objective tests. *Medical Teacher, 33* (6), 447-458. doi:10.3109/0142159X.2011.564682 [doi]

Tighe, J., McManus, I. C., Dewhurst, N. G., Chis, L., & Mucklow, J. (2010). The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: An analysis of MRCP(UK) examinations. *BMC Medical Education, 10*, 40-6920-10-40. doi:10.1186/1472-6920-10-40 [doi]

Veloski, J. J., Rabinowitz, H. K., Robeson, M. R., & Young, P. R. (1999). Patients don't present with five choices: An alternative to multiple-choice tests in assessing physicians' competence. *Academic Medicine: Journal of the Association of American Medical Colleges, 74* (5), 539-546.

Ward, W.C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement. 6* (1), 1–11.