



12-2014

Psychometrics and test-enhanced learning in a patient-centered learning curriculum

Syed Haris Ali

Follow this and additional works at: <https://commons.und.edu/theses>



Part of the [Philosophy Commons](#)

Recommended Citation

Ali, Syed Haris, "Psychometrics and test-enhanced learning in a patient-centered learning curriculum" (2014). *Theses and Dissertations*. 1074.
<https://commons.und.edu/theses/1074>

This Dissertation is brought to you for free and open access by the Theses, Dissertations, and Senior Projects at UND Scholarly Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UND Scholarly Commons. For more information, please contact zeineb.yousif@library.und.edu.

PSYCHOMETRICS AND TEST-ENHANCED LEARNING IN A
PATIENT-CENTERED LEARNING CURRICULUM

by

Syed Haris Ali
Bachelor of Medicine, Bachelor of Surgery, Sindh Medical College,
University of Karachi, 2006
Master of Science, University of Tromsø, Norway, 2009

A Dissertation
Submitted to the Graduate Faculty

of the

University of North Dakota

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Grand Forks, North Dakota
December
2014

UMI Number: 3681046

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3681046

Published by ProQuest LLC (2015). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Copyright 2014 Syed Haris Ali
ii

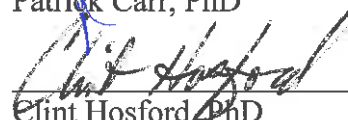
This dissertation, submitted by Syed Haris Ali in partial fulfillment of the requirements for the Degree of Doctor of Philosophy from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.



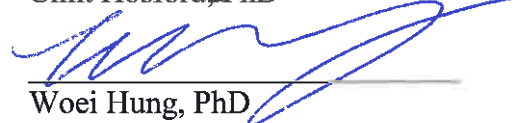
Kenneth Ruit, PhD



Patrick Carr, PhD



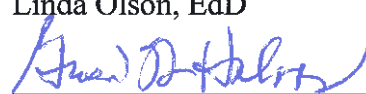
Clint Hosford, PhD



Woei Hung, PhD

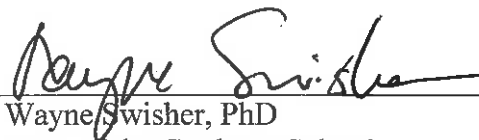


Linda Olson, EdD



Gwen Halaas, MD, MBA

This dissertation meets the standards for appearance, conforms to the style and format requirements of the Graduate School of the University of North Dakota, and is hereby approved.



Wayne Swisher, PhD
Dean of the Graduate School

September 24, 2014

Date

PERMISSION

Title: Psychometrics and Test-Enhanced Learning in a Patient-Centered Learning Curriculum

Department: Anatomy and Cell Biology

Degree: Doctor of Philosophy

In presenting this dissertation in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work, or in her absence, by the chairperson of the department or the dean of the Graduate School. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Syed Haris Ali
September 24, 2014

TABLE OF CONTENTS

LIST OF FIGURES	ix
LIST OF TABLES	x
ACKNOWLEDGMENTS	xii
ABSTRACT	xiii
CHAPTER	
I. INTRODUCTION	1
A Brief History of Patient-Centered Learning Curriculum	1
Components of a Curriculum	3
Introduction to Test-Enhanced Learning (Research Question 1)	6
Introduction to Assessment in Undergraduate Medical Education (Research Questions 2 and 3)	8
References	17
II. VALIDITY AND RELIABILITY OF SCORES OBTAINED ON MULTIPLE-CHOICE QUESTIONS: WHY FUNCTIONING DISTRACTORS MATTER	21
Abstract	21
Background	21
Methods	21
Results	22
Conclusion	22

Introduction.....	22
Validity of Scores Obtained on Multiple-choice Questions.....	22
Reliability of Scores Obtained on Multiple-choice Questions.....	30
Materials and Methods.....	33
Research Design.....	33
Subjects and Setting.....	33
Sample of Questions	34
Procedure	35
Intervention.....	35
Data Collection and Analysis.....	36
Results.....	38
Validity of Scores Obtained on Multiple-choice Questions.....	39
Reliability of Scores Obtained on Multiple-choice Questions.....	44
Discussion.....	49
References.....	57
III. THE IMPACT OF ITEM FLAWS, TESTING OF LOW COGNITIVE LEVEL, AND LOW DISTRACTOR FUNCTIONING ON MULTIPLE-CHOICE QUESTION QUALITY	61
Abstract.....	61
Background.....	61
Method	61
Result	62

	Conclusion	62
	Introduction.....	62
	Materials and Methods.....	70
	Research Design.....	70
	Subjects.....	70
	Procedure	71
	Intervention.....	71
	Data Collection and Analysis.....	74
	Results.....	75
	Experimental Subgroup A.....	75
	Experimental Subgroup B.....	77
	Control Group (C).....	78
	Discussion	81
	References.....	88
IV.	SHORT- AND LONG-TERM RETENTION OF KNOWLEDGE OF HUMAN ANATOMY – THE ROLE OF RETRIEVAL PRACTICE	91
	Abstract.....	91
	Introduction.....	91
	Materials and Methods.....	91
	Results.....	92
	Conclusions.....	93
	Introduction.....	93
	Materials and Methods.....	101

	Subjects	101
	Materials	102
	Procedure	102
	Outcome Analysis.....	104
	Results.....	105
	Short-term retention	105
	Long-term retention	106
	Discussion	109
	References.....	117
V.	SUMMARY AND CONCLUSION	119
	References.....	125

LIST OF FIGURES

Figure	Page
II-1. Example of Free-Response (FR) and Multiple-Choice (MCQ) versions of an item	35
II-2. Effect size (Cohen's d) of the difference between mean expected and observed MCQ difficulty indices	42
II-3. Percentage of MCQ distractors with different selection frequencies	44
II-4. Standard deviation and reliability coefficient (Cronbach's alpha) of scores obtained on multiple-choice version of the exam	47
IV-1. Examples of questions used in repeated testing via free-response (Subgroup A topics) and multiple-choice (Subgroup B topics).....	103

LIST OF TABLES

Table	Page
II-1. Calculation of expected MCQ difficulty index from FR version difficulty and number of MCQ options	28
II-2. Demographic characteristics of each cohort of subjects.....	34
II-3. Number of students taking the Free-Response (FR) and Multiple-Choice (MCQ) versions of the exam in all cohorts	39
II-4. Mean Free-Response (FR) and Multiple-Choice (MCQ) difficulty indices in all cohorts	40
II-5. Number of total MCQ distractors as well as the number of distractors with $\geq 5\%$, $\geq 10\%$, $\geq 20\%$ and $\geq 33\%$ selection frequency in each cohort.....	42
II-6. Number of students taking the FR and MCQ versions of the exam, mean score, standard deviation, reliability coefficient (Cronbach's alpha) and Standard Error of Measurement (SEM) in all cohorts	45
III-1. List of IWFs, published by NBME3, with corresponding numerical codes used in this study.....	65
III-2. Demographic characteristics in each cohort of subjects.....	70
III-3. Example of interventions in Experiment Subgroup A (Removal of IWFs and enhancement of tested cognitive level).....	72
III-4. Example of interventions in Experiment Subgroup B (Replacement of non-functioning distractors)	74
III-5. Flaw type, tested cognitive level (CL), # of functioning distractors (FDs), difficulty index (diff.) and point biserial correlation (pbi) before and after intervention in Experimental Subgroup A (IWF removal + enhancement of tested CL) of knowledge	76

III-6.	Flaw type, tested cognitive level (CL), # of functioning distractors (FDs), difficulty index (diff.) and point biserial correlation (pbi) before and after intervention in Experimental Subgroup A (IWF removal + enhancement of tested CL)	77
III-7.	Flaw type, tested cognitive level (CL), # of functioning distractors (FDs) and psychometric characteristics before and after intervention in Control group (C; no intervention)	79
III-8.	Summary of psychometric characteristics before and after intervention in experiment and control group, as well as the result of Fisher's exact analysis	79
IV-1.	Short-term retention (average in-lab item difficulty vs. end-of-block exam item difficulty) of radiologic anatomy topics	106
IV-2.	Short-term retention (average in-lab item difficulty vs. end-of-block exam item difficulty) of non-radiologic anatomy topics	107
IV-3.	Long-term retention (end-of-block exam vs. end-of-year quiz item difficulty) of radiologic anatomy topics	108
IV-4.	Long-term retention (end-of-block exam item difficulty vs. end-of-year quiz item difficulty) of non-radiologic anatomy topics	108

ACKNOWLEDGMENTS

The work herein is dedicated to my siblings, Kinza Haider and Wajid Ali.

Dr. Haider and Dr. Ali: thank you for everything!

I'd also like to thank my doctoral committee for their support throughout my doctoral education, especially my advisor Kenneth G. Ruit, PhD and close mentor, Linda Olson, EdD.

ABSTRACT

Validity and reliability of scores obtained on Multiple-choice questions (MCQs), as well as the benefits of test-enhanced learning, have been of interest to medical educator scholars. Presented in this dissertation are four composite studies on these themes. The following hypotheses were tested:

1. Increased MCQ distractor functioning increases the validity and reliability of obtained scores.
2. Correction of item writing flaws (along with enhancement of tested cognitive level) and replacement or removal of non-functioning distractors equally improves psychometric characteristics of MCQs.
3. Repeated testing via free-response items enhances the retention of knowledge of human anatomy, compared with repeated or once-testing via multiple-choice questions.

Validity and reliability of scores obtained on MCQs was noted to rise as a result of increased MCQ distractor functioning, discrimination amongst high and low performing students was found to be equally improved via removal of flaws and non-functioning distractors, and short-term (up to four weeks) retention of human anatomy knowledge was found to be enhanced by repeated testing via free-response questions. Raising MCQ quality by addressing flaws and low distractor functioning, and using no-stakes repeated retrieval practice, is advised for improvement in the assessment and learning practices in pre-clinical medical education.

CHAPTER I

INTRODUCTION

A Brief History of Patient-Centered Learning Curriculum

Curriculum is defined as all the learning, which is planned and guided by the school, whether it is carried on in groups or individually, inside or outside the school.¹ It is one of the components of the educational process; other elements include teachers, students, physical facilities such as audiovisual media, laboratories, libraries, etc. Historically, curricula in undergraduate (Years 1 – 4) medical education have been classified into a few major approaches²: apprenticeship model (1765– present), discipline-based model (1871–present), organ-system- based model (1951–present) and problem-based learning (PBL) model (1971–present). A brief description of the journey toward Problem-based Learning model follows.

The Carnegie Foundation for the Advancement of Teaching is a U.S.-based education policy and research center founded in 1905 by a renowned philanthropist Andrew Carnegie. In early 20th century, the Carnegie Foundation for the Advancement of Teaching commissioned a reputed educator, Abraham Flexner, to study and report on the state of medical education in North America. At that time, there were 155 medical schools in North America, which differed greatly in their curricula, methods of assessment, and requirements for admission and graduation. Flexner visited all 155 schools, gathered important information, and wrote a comprehensive report that was published by the Carnegie Foundation in 1910.³ The report entitled “Medical Education

in the United States and Canada'' made recommendations for emphasis on scientific basis of medicine, bachelor level premedical studies, and two years of basic medical science instruction followed by two years of supervised clinical experience during undergraduate medical education.^{2,3} These recommendations were taken up by various institutions across North America and led to designing of curricula along the lines of various basic medical science disciplines such as Anatomy, Physiology, Biochemistry, Pharmacology, Microbiology and Pathology. With the passage of time, a concern arose that discipline-based instruction may promote rote memorization of facts over conceptual learning, thereby causing a delay in students' abilities to associate basic science with real-patient situations.⁴ Therefore, early clinical exposure began to be emphasized in undergraduate medical education and resulted in the Problem-Based Learning (PBL) curricular model prevalent today.⁴

Problem-based Learning is based on educational application of cognitive sciences and clinical reasoning theories that are meant to guide the development of students from novices to experts.⁴ This model of learning makes use of carefully designed clinical problems that demand knowledge acquisition, problem solving ability, self-directed learning strategies, and team participation skills from the learner. Students work in small groups, generate hypotheses and learning objectives, accumulate knowledge individually in their own time, and then reconvene to teach each other and solve the clinical problem under discussion. Problem-based Learning provides learners an opportunity to more aptly apply acquired knowledge in new contexts and fosters an environment that encourages self-directed learning and team-working skills deemed to be important in medical practice.^{5,6}

While Problem-based Learning has gained increasing popularity over the last two decades with most U.S. medical schools adopting it to some extent, of late, a concern regarding its problem-focused nature has been raised. The concern is that it lacks the acknowledgement that a patient is more than her or his biology or symptoms.⁷ This concern has given rise to a hybrid form of Problem-based Learning termed Patient-Centered Learning (PCL). Patient-centered Learning uses social concerns as key aspects of the clinical encounter; it integrates the biology of disease with its psychosocial determinants, and is meant to foster greater communication and partnership amongst healthcare team members.⁸ Preclinical medical education curriculum at the University of North Dakota School of Medicine and Health Sciences is a hybrid of patient-centered learning as well as the traditional discipline-based instruction. The faculty behind its development published their thoughts in the journal *Academic Medicine* in the year 2007, noting that adaptation of Patient-centered Learning has brought a change in institutional culture through promotion of professionalism education and through provision for a forum that supports, models, and promotes relationship-centered professional values⁷.

Components of a Curriculum

Curricula of any type have three major components: content, educational or instructional methods and assessment.^{9, 10} A brief introduction to these components follows.

The first major component of any curriculum is “content”. The content of a curriculum comprises a selection of knowledge, skills, and attitudes relevant to a profession and forms the basis for learning objectives of a given educational experience.¹⁰ Content should reflect the tasks learners will be performing after

completion of their education and should be tailored to a level appropriate for the learners. Curricular content can be gathered from many sources. Such sources include existing curriculum at one's own institution, professional associations that define appropriate content in their relevant discipline, textbooks, and educational needs assessment.¹¹ An "educational needs assessment" uses input from a variety of stakeholders to develop and prioritize goals and objectives of that educational experience.¹⁰ The *goals* of an educational experience are supposed to be broad, while the *objectives* of an educational experience are supposed to be specific and measurable.¹⁰ A formula, created by Kern et al.,¹⁰ for writing appropriate educational goals and objectives is worthy of mention here. The formula is, "Who will do how much or how well of what by when."¹⁰ Using this formula, an example goal of an educational experience in cadaveric anatomy would be, "By the end of second curricular block, Year 1 medical students will describe anatomy of the thoracic wall and its cavity, and will apply the acquired knowledge to relevant clinical contexts". Within this broad goal, an example educational objective would be, "By the end of second curricular block, Year 1 medical students will identify various muscles inserting on, or originating from, the ribs". The goals and objectives of an educational experience should be delineated with appropriate attention to constraints in terms of contact hours, availability of instructors and access to learning resources.¹⁰

The second major component of any curriculum is "educational (or instructional) method". The choice of educational method is dependent on the individual instructor. However, it is expected that the instructor will choose an educational method that is appropriate for the type of learning objectives to be accomplished through that

educational experience.¹⁰ There are three major types of learning objectives: knowledge objectives, skill objectives and attitudinal objectives. In regards to knowledge objectives, suitable educational methods would be “lectures” and “laboratory experiences.”¹⁰ In regards to skill objectives, suitable educational methods would be “simulation exercises” and “standardized patient encounters.”¹⁰ In regards to attitudinal objectives, a suitable educational method would be “group discussions” or “self-assessment essays.”¹⁰ Problem-based learning is also a type of educational method that helps accomplish different types of educational objectives, such as knowledge and attitudinal objectives, simultaneously; however, prescribed guidelines should be followed by the participants of problem-based learning sessions in order for learning objectives to be met in a systematic and organized manner.^{5-7, 10} Regardless of the choice of educational method, it is important to ensure that methods are appropriately matched with the learning objectives and that educational methods facilitate the attainment of stated objectives in a clear and explicit manner. Moreover, there should be a clear mechanism to support learners in regards to self-directed learning; such support is usually offered via counseling on study strategies and adequate study resources.^{10, 11}

Recent technological advances have given rise to a variety of new educational methods. Methods such as “online” learning have a room for flexibility and customization that benefit the instructor and students alike.^{10, 11} Moreover, the role of repeated, no-stakes assessment as an educational method has also been discussed in medical education literature.¹² Since one of the chapters in this dissertation is on this very topic, a brief introduction to the concept of testing-enhanced learning follows.

Introduction to Test-Enhanced Learning (Research Question 1)

The main purposes of assessment are to assign grades, certify competence and evaluate curricular effectiveness in terms of meeting curricular goals. Assessment in medical education has undergone frequent revisions over the past several decades owing to how clinical “competence” is conceptualized.¹³ Both the *content* and *format* of assessment has been of interest to educational psychologists and medical educator scholars. In regards to the format of assessment, the multiple-choice question, an old and trusted tool for educational assessment, is favored extensively to this day for its quicker and objective scoring.¹⁴ However, when concerns were raised that multiple-choice questions tend to assess plain recall and not complex thought process,¹⁵ the focus shifted to open-ended, modified essay questions to help assess problem solving ability.^{16, 17} Another argument was made that problem solving is based on the fund of knowledge and that evaluation of memory constructs versus recall of such knowledge should allow better judgment of learners’ competence.¹⁶ This argument led to development of alternate formats of assessment. Such formats include “key features” problems, which emphasize concepts necessary for problem solving, and “script-concordance questionnaires”, which compare knowledge organization between novices and experts.^{13, 14} These assessment tools are being used in assessment in undergraduate and postgraduate medical education to varying degrees. However, multiple-choice questions have remained the mainstay of assessment in undergraduate medical education owing to a high correlation between scores obtained on multiple-choice questions and their open-ended, free-response counterparts.^{18, 19} This shows that in the contemporary tradition in undergraduate medical education, *content* of assessment is emphasized over the *format* of assessment.^{14–19}

The educational benefit of multiple-choice assessment has also been dwelled upon in literature.^{15,20} The oft-repeated saying is that “assessment drives learning”, and frequent testing has been found to encourage deep learning and modification of learning strategies.²¹ Moreover, frequent testing, combined with input from learners, has been reported to create opportunities for modification of educational methods and overall curricular improvement as well.²²

Frequent or repeated testing via both free-response and multiple-choice formats has been found to be useful in enhancement of learning.¹² Tests comprising free-response questions are also termed “production tests”, since this format of testing requires production of the answer from memory.¹² On the other hand, tests comprising multiple-choice questions are also termed “recognition tests”, since this format of testing requires recognition of the answer from a list of options.¹² It has been suggested that knowledge is retained to a greater degree when educational content is tested repeatedly via free-response questions (production tests) than via multiple-choice questions (recognition tests).²³ Cognitive psychology explains this phenomenon via the amount of effort needed to recall the information; free response questions require more forceful retrieval of information than multiple-choice questions.²³ This allows enhancement of neuronal connections pertaining to a memory leading to better and longer degree of knowledge retention.²⁴ Therefore, repeated practice of information retrieval via free response questions has been argued to knowledge retention to a greater degree.²⁴

The last chapter (Chapter 4) of this dissertation is based on a study of the value of test-enhanced learning in the context of undergraduate medical education in human anatomy. The purpose of the study was to evaluate the usefulness of no-stakes testing as

an educational method in undergraduate medical education and to replicate the findings of others in the context of learning of human anatomy. The research question of that study is, “does the format (free-response vs. multiple-choice) and frequency (repeated vs. once) of tests using different questions on the same topic influence short- (4 weeks) and long-term (2 – 7 months) retention of anatomy knowledge? In that chapter, the concept of test-enhanced learning and findings from other studies are discussed in greater detail. Also, results from our yearlong investigation are presented and discussed in regards to the outcomes of interest for the educators and the learners.

Introduction to Assessment in Undergraduate Medical Education (Research Questions 2 and 3)

The third major component of any curriculum is “assessment”. As discussed above, assessment in undergraduate medical education is heavily reliant on Multiple-choice Questions (MCQs).^{9, 10, 13} Multiple-choice questions consist of a stem (the “question” statement), a few incorrect options (the “distractors” or “foils”), and one correct option.²⁵ High-quality multiple-choice questions afford the advantage of assessing large numbers of students with efficiency and objectivity²⁵ and usually provide sufficient discrimination amongst high and low ability students.²⁶ However, quality of multiple-choice questions is very important for valid (accurate) assessment of learners’ knowledge.^{25, 26} Moreover, the tendency of poorly constructed multiple-choice questions to assess factual recall (instead of higher order thinking) has been reported to be a significant impediment to accurate assessment of student knowledge.^{27, 28}

Two issues with poorly constructed multiple-choice questions are, a. lack of functioning distractors, and b. item writing flaws.²⁹⁻³¹ A functioning distractor is an incorrect option that is selected by $\geq 5\%$ of examinees ($\geq 5\%$ selection frequency).^{25, 29}

Another property desirable in a functioning distractor is that it should be chosen more by low-performing examinees than high-performing examinees.^{25, 29} Such selective attractiveness for low-performing students renders “negative” discriminatory ability, which is a desired trait in a functioning distractor.^{25, 29} Item writing flaws, on the other hand, are violations of accepted item-writing guidelines.^{30, 31} Lack of functioning distractors in a multiple-choice question allows test-wise students to guess the answer correctly without having the necessary prerequisite knowledge.²⁹ On the other hand, item writing flaws such as a “confusing stem” or “long options” increase the difficulty of an item; such difficulty may have no relevance to an item’s inherent difficulty. Increased difficulty of an item stemming of item writing flaws introduces a variance in performance, termed “construct-irrelevant variance.”³² And, construct-irrelevance variant has been reported to adversely impact the validity of scores obtained on multiple-choice questions.³²

Another issue discussed in the context of item writing flaws is testing of low cognitive function by flawed multiple-choice questions. A question assessing “low” cognitive level tends to assess plain recall of a fact, while a question assessing “high” cognitive level tends to assess application of factual knowledge to relevant contexts.²⁹ The concern with testing of low cognitive level is that competent learners, especially in the field of medicine, are expected to process complex information and make sound clinical decisions based on their fund of knowledge.¹⁴ Testing of low cognitive function (plain factual recall) precludes the assessment of such competence.³⁰ Assessment of low cognitive function and the prevalence of item writing flaws has been lamented upon in

the literature; two seminal and often cited studies show how flawed multiple-choice questions impede examiners' ability to gauge student knowledge accurately.^{30, 31}

The third chapter of this dissertation is based on a study of the value of addressing item flaws and enhancing the number of functioning distractors in high-stakes multiple-choice assessment in Year 1 medical education. The purpose of the study was to evaluate the usefulness of addressing these two issues in enhancing the evidence of validity of scores obtained on high-stakes in-house exams. The research question of that study was, "To what extent does correction of item flaws, and enhancement of tested cognitive level and number of functioning distractors impacts the difficulty and discriminatory ability of multiple-choice questions?" In that chapter, the concepts of item flaws, distractor functioning and cognitive level of assessment, as well as findings from other studies, are discussed in greater detail. Also, results from our investigation are presented and discussed in regards to the outcomes of interest for the basic medical science faculty that writes multiple-choice questions used in high-stakes in house assessment. What follows is a synopsis of the concepts of validity and reliability of multiple-choice assessment, since these two concepts form the basis of rest of the work presented in this dissertation.

In choosing the appropriate method to assess student learning, the starting point is the learning objectives of the curriculum or educational experience. Assessment must systematically analyze whether the learning objectives stated at the outset have been accomplished.¹¹ Choice of the appropriate method depends on the type of learning objectives under assessment. Multiple-choice or free-response type questions are better suited for assessment of knowledge type objectives.¹¹ On the other hand, standardized patient encounters and simulation exercises suit the assessment of skills, while self-

reflective essays suit the assessment of learning objectives pertaining to attitudes.¹¹ An examination blueprint is helpful in mapping out and ensuring the coverage of essential learning objectives.⁹ Assessment should be designed to allow appropriate discrimination among high- and low-performing students.³³ Also, the rules and regulations governing assessment procedures must be clear and should be applied appropriately and consistently.^{9, 33}

The domain of *psychometrics* encompasses assessment in fields of psychology and education.³⁴ The concepts of *validity* and *reliability* fall under the domain of psychometrics; evidence of validity and reliability of obtained scores serves as an indicator of an exam's quality.⁹ Majority of the work presented in this dissertation deals with validity and reliability of scores obtained on the staple assessment tool in undergraduate medical education: the multiple-choice question. An introduction to these concepts follows.

The term *validity* refers to the degree to which conclusions derived from the results of any assessment are well grounded, justifiable and meaningful.^{34, 35} In other words, it describes how much the interpretation of a test result can be trusted. Many physical instruments measure a quantity. Examples of such quantities are height, blood pressure and plasma sodium level. Interpreting the meaning of result of such physical measurement is straightforward,³⁶ plasma sodium level greater than 145 milli-equivalents per liter indicate greater than normal level of sodium circulating in the body. In contrast, assessment of student knowledge, patient symptoms or physician attitude may yield results that are somewhat open to interpretation. Such assessments, in contrast to physical measurements described above, measure an underlying “construct”, which is

defined as “an abstract concept or principle.”³⁴ Assessment of student knowledge yields scores that have meaning specific to the construct of the assessment questions; that meaning reflects the validity of obtained scores.³⁴ It is important to note that validity is not a property of the method or instrument of assessment, but of the scores obtained on that assessment instrument.³⁷ For example, we would expect scores obtained on a board exam in pulmonology to accurately reflect knowledge of the construct “pulmonology”. However, scores obtained on such an exam would not accurately reflect knowledge of the constructs “cardiology” or “thoracic surgery”, even though the latter two constructs are somewhat related to the intended purpose of assessment, i.e. knowledge of pulmonology. Therefore, scores must be interpreted in light of only the intended construct of assessment.³⁴

Validity of scores obtained on an assessment method or tool (such as multiple-choice questions) must be established through evidence. Such evidence is gathered from various sources.^{32, 34} The sources of evidence of validity help examiners in making definite conclusions regarding student achievement. For example, consider an exam of knowledge of histology given to Year 1 medical students. Now consider that that evidence was gathered in regards to the validity of scores obtained on that exam, and that evidence was found to be strong. This would mean that scores obtained on that particular exam are valid for the constructs (concepts or topics) the exam questions purported to assess. Now consider that 25% of the examinees failed that exam. Since the scores obtained on that exam are being considered valid (owing the strength of validity evidence), examiners can confidently conclude that students who failed that exam truly

did not possess the requisite knowledge of histology at the time of the exam. Such a conclusion is dependent upon the strength and sources of validity evidence.³⁴

Five sources of evidence of validity have been defined in contemporary psychometrics literature.^{34,37} These sources are “content”, “response process”, “internal structure”, “relations to other variables”, and “consequences.”^{34,37} The following descriptions of these sources of validity evidence are based on the work published by Messick³⁴ and Downing.³⁷ The “content” source of validity evidence assesses how well assessment questions (items) represent the intended construct. The “response process” source of evidence assesses the strength of relationship between the intended construct and the thought processes of examinees. The “internal structure” source of validity evidence assesses psychometric characteristics of the assessment items; an extensive discussion of psychometric characteristics of assessment items lies ahead in this dissertation. The “relations to other variables” source of evidence assesses closeness of scores obtained on one assessment instrument with scores obtained on a reference instrument for that type of learning objective. The “consequences” source of validity evidence assess whether scoring high or low on an assessment method really makes a difference in practical terms. A combination of evidence from several different sources is necessary to support any given interpretation of obtained scores, and strong evidence from one source does not obviate the need for other supporting evidence.

The work presented in the first two chapters of this dissertation is based on an investigation of the validity of scores obtained on in-house multiple-choice exams. Specifically, the investigation focuses on two sources of validity evidence: relations to

other variables (Chapter 1), and internal structure (Chapter 2). These two sources will be discussed in greater detail in their respective chapters.

Reliability of psychometric instruments refers to the reproducibility or consistency of scores from one assessment to another.³⁸ The concept of reliability is interconnected with the concept of validity; an instrument that does not yield reliable scores does not permit its valid interpretation either.^{32,37} For example, blood pressure readings of 160/90 mmHg, 80/40 mmHg, and 140/60 mmHg over a few minutes' period in a stable patient would be considered unreliable. Scores obtained on educational assessment instruments are just as susceptible to unreliability. However, there is something peculiar about educational assessment instruments: it is often impractical or even impossible to administer the same exam to an individual or group of students twice or multiple times. Thus, it is important to accumulate evidence to establish the reliability of scores before using an assessment instrument in practice.

There are numerous ways to categorize and measure reliability.³⁸ The usefulness of various methods varies according to the assessment instrument. Assessments of reliability over time (test-retest) and between raters (inter-rater) are some of the commonly used methods in the field of education.³² Generalizability theory is another well-known method that provides a unifying framework for the various methods of reliability assessment.³² Under the framework of Generalizability Theory, the unreliability of scores is attributed to various sources of error such as items, raters and subjects. A variance in performance arises from such error sources. In the application of Generalizability Theory, the contribution of each error source is quantified systematically through complex statistical analysis.³⁹

The reliability of scores obtained on multiple-choice exams is assessed via the concept of “internal consistency.”³⁸ One of the commonly used measures of internal-consistency reliability is Cronbach’s alpha coefficient.³² The statistical derivation of this coefficient is based on the test-retest concept. The test-retest concept means that an exam is given on one occasion to a group of examinees, and the same exam (or an equivalent form of the same exam) is given to the same group of examinees at a later time (assuming that the examinees have not gained or lost any knowledge between the two administrations of the exam). If the exam produces reliable scores, the students should obtain nearly the same scores on second administration as on the first administration.³² While the test-retest concept is the foundation of internal-consistency reliability, the actual test-retest design is not used in actual practice since it is logistically impossible to give the same exam to the same group of students more than once. Therefore, to assess internal-consistency reliability via Cronbach’s alpha coefficient, the test-retest design is used to hypothetically divide an exam into two random halves (e.g., even-numbered items as the first half and the odd-numbered items as the second half). It is assumed that underlying constructs of the exam items are related to each other, and that two random halves are a reasonable proxy for two complete tests administered to the same group of examinees.³² Correlations of scores on all possible random halves of exam are calculated, and then averaged to generate the Cronbach’s alpha coefficient.³² A Cronbach’s alpha coefficient of at least 0.8 is considered desirable for high-stakes in-house multiple-choice exams.^{32, 38, 39}

To summarize, validity of scores obtained on an exam is based on evidence gathered to support proposed interpretations of results. Reliability pertains to the

reproducibility of scores across several administration of an exam. The research question of the study presented in the first chapter of this dissertation is, “what is the role of functioning distractors in validity and reliability of scores obtained on a multiple-choice exam of neurohistology knowledge?” This question arises from the need of a clear understanding of validity and reliability among medical educator scholars. The findings reported in this dissertation may advance the understanding of factors affecting the quality of multiple-choice assessment. Appropriate attention to such factors will help improve the quality of assessment in undergraduate medical education.

References

1. Kerr J. The problem with curriculum reform. In: Kerr J, editor. Changing the curriculum. London: University of London Press; 1968. p. 13–38.
2. Papa FJ, Harasym PH. Medical curriculum reform in North America, 1765 to the present: A cognitive science perspective. *Acad Med.* 1999;74(2):154–164.
3. Flexner A. A. Medical education in the United States and Canada. A report to the Carnegie Foundation for the advancement of teaching. Boston: Updyke; 1910.
4. Kinkade S. A snapshot of the status of problem-based learning in U.S. medical schools, 2003–04. *Acad Med.* 2005;80(3):300–301.
5. Whitfield CF, Mauger EA, Zwicker J, Lehman EB. Differences between students in problem-based and lecture-based curricula measured by clerkship performance ratings at the beginning of the third year. *Teach Learn Med.* 2002;14(4):211–217.
6. Norman GR, Schmidt HG. The psychological basis of problem-based learning: A review of the evidence. *Acad Med.* 1992;67(9):557–65.
7. Christianson CE, McBride RR, Vari RC, Olson L, Wilson HD. From traditional to patient-centered learning: Curriculum change as an intervention for changing institutional culture and promoting professionalism in undergraduate medical education. *Acad Med.* 2007;82:1079–1088.
8. Satterfield JM, Mittenness LS, Tervalon M, Adler N. Integrating the social and behavior sciences in an undergraduate medical curriculum: The UCSF essential core. *Acad Med.* 2004;79:6–15.
9. Harden RM. AMEE medical education guide No 21: Curriculum mapping: a tool for transparent and authentic teaching and learning. *Med Teach.* 2001;23:123–137.

10. Kern DE, Thomas PA, Hughes MT. Curriculum development for medical education: A six-step approach, 2nd edit. Baltimore, MD: Johns Hopkins University Press; 2009.
11. Harden RM. The integration ladder: a tool for curriculum planning and evaluation, *Med Educ.* 2000;34(7):551–557.
12. Larsen DP, Butler AC, Roediger HL. Test-enhanced learning in medical education. *Med Educ.* 2008;42:959–966.
13. van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ.* 1996;1:41–67.
14. Schuwirth LWT, van der Vleuten CPM. Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ.* 2004;38:974–979.
15. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ.* 1983;17:165–171.
16. van der Vleuten CPM, Newble DI. How can we test clinical reasoning? *Lancet.* 1995;345:1032–1034.
17. Palmer EJ, Devitt PG. Assessment of higher-order cognitive skills in undergraduate education: modified essay or multiple-choice questions? Research paper. *BMC Med Educ.* 2007;7:49.
18. Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Appl Psychol Meas.* 1982;6(1):1–11.
19. Schuwirth LWT, van der Vleuten CPM, Donkers HHL. A closer look at cueing effects in multiple-choice questions. *Med Educ.* 1996;30:44–9.

20. Schuwirth LWT, van der Vleuten CPM. Changing education, changing assessment, changing research? *Med Educ.* 2004;38:805–812.
21. Verhoeven BH, Verwijnen GM, Scherpbier AJJA, van der Vleuten CPM. Growth of medical knowledge. *Med Educ.* 2002;36:711–717.
22. Ringsted C, Henriksen AH, Skaarup AM, van der Vleuten CPM. Educational impact of in-training assessment (ITA) in postgraduate medical education: a qualitative study of an ITA programme in actual practice. *Med Educ.* 2004;38:767–77.
23. Butler AC, Roediger HL III. Testing improves longterm retention in a simulated classroom setting. *Eur J Cogn Psychol.* 2007;19:514–527.
24. Roediger HL III, Karpicke JD. The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci.* 2006;1:181– 210.
25. McCoubrie, P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach.* 2004;26(8):709–712.
26. Schuwirth, L.W.T., van der Vleuten, C.P.M. ABC of learning and teaching in medicine: written assessment. *Brit Med J.* 2003;326(7390):643–645.
27. Farley, J.K. The multiple-choice test: Developing the test blueprint. *Nurs Educ.* 1989;14(5):3–5.
28. Pampllett, R., Farnhill, D. Effect of anxiety on performance in multiple-choice examinations. *Med Educ.* 1995;29:298–302.
29. Downing, S.M. Assessment of knowledge with written test forms. In: Norman, G.R., Van der Vleuten, C., Newble, D.I. editors. *International Handbook of Research in Medical Education.* Dordrecht: Kluwer Academic Publishers; 2002. p. 647–672.

30. Downing, S.M. 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. *Adv Health Sci Edu.* 2005;10(2):133–143.
31. Jozefowicz, R.F., Koeppen, B.M., Case, S., Galbraith, R., Swanson, D., Glew, R.H. The quality of in-house medical school examinations. *Acad Med.* 2002;77(2), 156–161.
32. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am J Med.* 2006;119(166):167–116.
33. Tavakol M, Dennick R. Post examination analysis of objective tests. *Med Teach.* 2011;33:447–458.
34. Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Measure Issues Prac.* 1995;14:5–8.
35. Foster SL, Cone JD. Validity issues in clinical assessment. *Psychol Assessment.* 1995;7:248–260.
36. Bland JM, Altman DG. Statistics notes: validating scales and indexes. *Brit Med J.* 2002;324:606–607.
37. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ.* 2003;37:830–837.
38. Downing SM. Reliability: On the reproducibility of assessment data. *Med Educ.* 2004;38:1006–12.
39. Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No 66. *Med Teach.* 2012;34:e161–e175.

CHAPTER II

VALIDITY AND RELIABILITY OF SCORES OBTAINED ON MULTIPLE-CHOICE QUESTIONS: WHY FUNCTIONING DISTRACTORS MATTER

Abstract

Background

Plausible distractors (incorrect options) are important for accurate measurement of knowledge via multiple-choice questions (MCQs). This study demonstrates the impact of distractor functioning on validity and reliability of scores obtained on the MCQ version of an exam compared to its Free-response (FR) version.

Methods

FR and MCQ versions of a Neurohistology exam were randomly distributed among four different cohorts of Year 1 medical students. Distractor functioning as well as the disparity between item difficulty indices in the two exam versions was noted. Then, revision of the MCQ version of the exam was performed via replacement of consistently non-functioning distractors in with those developed from incorrect responses on FR version of the items. The revised MCQ version was given to all students in Cohort 5 as well as to random half of Cohort 6, whose other half received the FR version. Validity of MCQ scores was assessed by comparing an index of *expected* MCQ difficulty (calculated from the FR difficulty index) with the index of *observed* MCQ difficulty before and after replacement of previously non-functioning distractors. Reliability of

MCQ scores was assessed via calculation of Cronbach's alpha coefficient before and after replacement of previously non-functioning distractors.

Results

An increase in the number of MCQ distractors with $\geq 5\%$, $\geq 10\%$, $\geq 20\%$ and $\geq 33\%$ selection frequency was noted after replacement of previously non-functioning distractors with those developed from incorrect responses on the items' FR version. As a result of increased distractor functioning in Cohort 6, the difference between mean *expected* and *observed* MCQ difficulty indices was noted to be reduced thereby strengthening the evidence of validity of MCQ scores. Cronbach's alpha coefficient of MCQ scores was also noted to be higher after replacement of previously non-functioning distractors highlighting the improvement in internal consistency reliability of obtained scores.

Conclusion

Replacement of previously non-functioning MCQ distractors with those developed from incorrect responses on free response version of the items is helpful in enhancing the validity and reliability of MCQ scores.

Introduction

Validity of Scores Obtained on Multiple-choice Questions

The first part of the study presented in this chapter deals with the conceptual framework of validity. Validity is defined as the extent to which scores obtained on an assessment instrument represent true knowledge.¹ To assess an exam's ability to elicit true knowledge, systematic collection of validity evidence of exam scores is advised.² One of the sources 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.² 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 format.² A typical free-response (FR) item is a fill-in-the-blank question in which no options are provided and examinees have to write or type their answer.^{2,3} An un-cued (UnQ) item is a type of free-response item; the only difference between free-response and un-cued formats is that the examinee writes or types the answer's code and not the answer itself.⁴ The code is listed next to the answer in a booklet that contains hundreds of possible answers for all items in an exam; this booklet is provided only during the exam to all examinees.⁴ Both formats (free-response and un-cued) require production of the answer from examinee's memory. Therefore, these formats are also termed *production* tests.⁴ Also, these formats are devoid of the guessing and cueing inherent in multiple-choice questions, and therefore serve as a yardstick for evaluating the quality of multiple-choice version of an item.^{3,4} Studies comparing performance on exams written in multiple-choice and free-response or un-cued formats can be found in the literature.^{4,7} Findings from these studies are discussed below.

The study by Damjanov et al. was based on the material covered in Year 1 medical education.⁴ In alternate years, questions in the subject of pathology were given in the standard multiple-choice format or in the open-ended un-cued (UnQ) format. The study found no significant difference between students' mean scores or item discrimination indices on the two versions of the exam. Damjanov et al. concluded that the un-cued open-ended format of assessment is an acceptable alternative to the multiple-choice format. The study by Fajardo et al. compared performance on un-cued and

multiple-choice formats of various items given in a summative evaluation in a radiology clerkship.⁵ Students' level of performance was found to be lower on un-cued version of the items (mean score = 68.9 ± 10.2 standard deviation) than on multiple-choice version of the items (mean score = 75.6 ± 12.4 standard deviation). The study showed how un-cued version of an exam helps in assessment of recall of critical information without the threat of guessing or cueing inherent in multiple-choice testing. Like Damjanov et al., Fajardo et al. also suggested that un-cued format of items helps in overcoming some of the limitations of conventional multiple-choice testing. The study by Prihoda et al. proposed a “correction for random guessing” on multiple-choice questions given in an oral and maxillofacial pathology exam to Year 2 dental students.⁶ The correction was a weighting formula for points awarded for correct answers, incorrect answers, and unanswered questions such that the expected value of the increase in test score due to guessing was zero. Uncorrected and corrected scores were compared with the free-response counterpart of the multiple-choice exam since free-response format greatly reduces the potential for correct guessing. It was found that the agreement between corrected multiple-choice scores and free-response scores was greater (intraclass correlation coefficient 0.78, $p = 0.015$) than the agreement between uncorrected multiple-choice scores and free-response scores (intraclass correlation coefficient 0.71) thereby highlight the value of correction for guessing in enhancing the validity of obtained scores. Prihoda et al. concluded that correction for guessing renders higher validity to multiple-choice scores and that examiners should be wary of guessing and cueing inherent in multiple-choice questions before deriving definite conclusions from obtained results.⁶ The study by Newble et al. compared performance of medical students and practicing

physicians on a test of clinical knowledge written in multiple-choice and free-response formats.⁷ Scores were found to be generally higher on multiple-choice version of the test than on the free-response version. However, the difference between mean scores obtained on the two exam versions was found to be smaller among practicing physicians (19% difference) than among senior-level (37% difference) and junior-level (61% difference) students. Newble et al. surmised that students performed much better on the multiple-choice version than on the free-response version due to guessing and cueing afforded by the multiple-choice questions. On the other hand, practicing physicians performed similarly on both versions of the exam owing to their fund of knowledge (expertise) and lesser reliance on guessing and cueing.⁷ A questionnaire survey given in this study showed that students were aware of the deficiencies in multiple-choice testing, and a large majority believed that free-response testing gave a more accurate assessment of their clinical ability. Newble et al. concluded that in tests aimed at measuring clinical competence, multiple-choice questions appear to overestimate examinees' ability, which makes them less suitable than free-response questions for assessment of clinical competence.

The difference between performance on an item's free-response and multiple-choice versions is mainly attributed to the functioning of its distractors.^{3,8} A *functioning* distractor (FD) is an incorrect option that is selected by $\geq 5\%$ of examinees (i.e., $\geq 5\%$ selection frequency).⁸ Another property desirable in a functioning distractor is that it should be chosen more by low-performing examinees than high-performing examinees.⁸ Such selective attractiveness for low-performing students renders "negative" discriminatory ability to that distractor, which is a desired trait in a functioning

distractor.⁸ On the other hand, a *non-functioning* distractor (NFD) is an incorrect option chosen by fewer than 5% examinees and possesses a positive discriminatory ability, both of which are undesirable characteristics in a multiple-choice distractor.⁸ Tarrant et al. have reported on the impact of eliminating a non-functioning distractor from a 4- or 5-option multiple-choice question.⁹ The aim of their study was to study the effect of such removal on psychometric properties (difficulty and discriminatory ability) of multiple-choice questions. Using item-analysis data, they eliminated the distractor with the lowest selection frequency and compared performance on the 3- and 4-option versions of 41 multiple-choice questions in two cohorts of nursing students. They found that removing the non-functioning distractor resulted in minimal change in mean item difficulty (0.3%). The three-option version of the items were found to contain more functioning distractors despite having fewer distractors overall. Moreover, existing distractors were found to be more discriminatory when infrequently selected distractors were removed from the multiple-choice questions. Since three-option questions require less time to develop and administer, Tarrant et al. encouraged adoption of three-option multiple-choice questions as the standard in multiple-choice testing.⁹

From the study by Tarrant et al.,⁹ it becomes obvious that a 5-option version of a multiple-choice question is not superior to its 4-option or 3-option version, if the 4th and 5th options lack plausibility. A seminal study published by Alex Rodriguez based on 80 years of research in this area agrees with this notion.⁸ Consolidating the findings from dozens of previously published studies, Rodriguez's meta-analysis showed that systematically removing one non-functioning distractor from 5-option multiple-choice questions reduced their average difficulty and discriminatory ability only to a mild extent

(0.02 and 0.04 units, respectively). Moreover, removing two non-functioning distractors from 5-option multiple-choice questions caused average item difficulty to reduce a little further (0.07 units), with no effect on average item discriminatory ability. The studies by Tarrant et al.⁹ and Rodriguez⁸ show that non-functioning distractors offer very little in terms of difficulty and discriminatory ability of multiple-choice questions. Also, both authors argue that removal of non-functioning distractors may reduce response time needed per multiple-choice question. This allows inclusion of more multiple-choice questions in an exam of fixed duration, which lends the benefit of increased content sampling for that exam.^{3, 8, 9} However, one must note that these benefits are the outcome of removal of only *non-functioning* distractors; these studies highlight the important role of plausible, *functioning* distractors in accurate assessment of knowledge via multiple-choice questions.^{3, 8, 9}

The work presented in this chapter of the dissertation advances our understanding of the role of distractor functioning in rendering validity and reliability to scores obtained on multiple-choice questions. Two versions (free response and multiple choice) of the same Neurohistology exam were randomly distributed among four different cohorts of Year 1 medical students. We made a note of distractor functioning on multiple-choice questions, as well as the disparity between difficulty indices on free-response and multiple-choice versions of the items. Then, the multiple-choice version of the exam underwent revision via replacement of consistently non-functioning distractors with those developed from incorrect responses on free response version of the items. The revised multiple choice version of the exam was given to all students in Cohort 5 and to random half of Cohort 6, whose other half received the free response version of the exam.

Evidence of validity of scores obtained on the multiple-choice version of the exam was collected before and after this revision. The collected evidence pertained to the source “Relations to other variables” that was described in the beginning of this chapter.^{2,3} Collection of this evidence entailed calculation of an index of *expected* MCQ difficulty, and its comparison with the index of *observed* MCQ difficulty. The index of *expected* MCQ difficulty was calculated by postulating that a certain proportion of students who answered an item incorrectly on its free response version would have answered the item correctly on its multiple choice version through random guessing. An example of calculation of the index of *expected* MCQ difficulty is as follows.

Suppose, the free response version of an item is correctly answered by 60% of examinees (FR difficulty index: 0.6). The proportion of students with an incorrect answer on the free response version would be 40% [0.4]. Now suppose that multiple-choice version of this item contained 5 options. It will be anticipated that a certain proportion of students who answered the item incorrectly on its free response version might have been able to guess it correctly on its multiple-choice version using random guessing among the 5 options. Probability would suggest that such a proportion among 40% (0.4) of students would be 8% (0.08) ($0.4 / 5 = 0.08$). The addition of that proportion of students (0.08) to the free-response difficulty index will yield the index of *expected* MCQ difficulty ($0.6 + 0.08 = 0.68$). This example is laid out in Table II-1.

Table II-1. Calculation of expected MCQ difficulty index from FR version difficulty and number of MCQ options.

MCQ ID	# of total options	FR version difficulty (FR diff.)	Proportion of students with incorrect answers on FR version (P_w)	Expected inflation in item ease (EI) ($P_w / \#$ of total options in the MCQ version)	Expected MCQ difficulty (FR diff. + EI)
Example	5	0.60	$1 - 0.60 = \mathbf{0.40}$	$0.40 / 5 = \mathbf{0.08}$	$0.60 + 0.08 = \mathbf{0.68}$

The comparison between *expected* and *observed* MCQ difficulty indices was made before and after replacement of previously non-functioning distractors. However, the comparison was based on two assumptions. The assumptions were, a. the free response version of the item elicits true knowledge, and b. faculty responsible for the assessment of basic science content writes reasonably plausible multiple-choice distractors. The basis for the first assumption is well established in the published literature and comes from the fact that free-response testing involves minimal guessing and cueing.³⁻⁷ The basis for the second assumption is that faculty responsible for teaching and assessing basic science content is well placed to write plausible multiple-choice distractors owing to their subject matter expertise.

Both versions (free-response and multiple-choice) of the exam were randomly distributed among each cohort of students to prevent selection bias and allow adequate comparison of actual multiple-choice performance (*observed* difficulty index) with what it ought to have been (*expected* difficulty index). Availability of performance data on the free response version of the items was pivotal in this regard. To date, no such comparisons of *expected* and *observed* MCQ difficulty indices have been reported in the context of assessment in undergraduate medical education, which highlights the novelty of presented study. Research hypothesis of this part of the study (validity of scores obtained on multiple-choice questions) 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.

Reliability of Scores Obtained on Multiple-choice Questions

The second part of the study presented in this chapter deals with the conceptual framework of reliability. The concept of reliability is related to Classical Test Theory,¹⁰ the central tenet of which is that an examinee's score (X) can be decomposed into her/his true score (T) and a random error component (E) ($X = T + E$). An examinee's true score (T) is defined as the score obtained if the exam was measuring the ability of interest perfectly (i.e. with no measurement error). A reliability coefficient, which ranges from 0 to 1, estimates of the level of concordance between observed and true scores of an examinee.¹⁰ It can also be interpreted as the proportion of variance among scores obtained by different examinees explained by the difference among abilities of those examinees.¹⁰ The proportion of variance among examinee scores explained by factors other than the difference among examinees' abilities is an unwanted phenomenon, and is classified as random error.¹⁰

A value of zero (0) reliability coefficient means no concordance (all error), whereas a value of one (1) means perfect concordance (all variance attributable to examinee abilities) between observed and true scores of the examinees.¹⁰ The reliability measure of interest in this chapter of the dissertation is *internal consistency* reliability. Internal consistency reliability was discussed in detail in the introductory chapter of this dissertation and, simply put, is the measure of reliability in exams that require a single administration (i.e., exams that cannot be given multiple times to the same group of examinees).¹¹ A coefficient of *internal consistency* reliability depicts 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.¹¹ One such coefficient is Cronbach's alpha; for high stakes assessment such as in-house curricular block exams, Cronbach's alpha coefficient of at least 0.8 is desired.^{10, 11} Derivation of the Cronbach's alpha coefficient is discussed in more detail in the Methods section of this chapter. Findings from a few studies on reliability in multiple-choice exams in undergraduate and postgraduate medical education are discussed below.

The study by Damjanov et al. was based on pathology content covered in pre-clinical medical education⁴. In alternate years, the questions were presented in the standard multiple-choice format or open-ended, un-cued (UnQ) format. Reliability coefficient of scores obtained on multiple-choice version of the exam was found to be 0.53, while that on un-cued version of the exam was found to be 0.63.⁴ It was concluded that scores obtained on un-cued items tend to be more reliable than their multiple-choice counterparts. The study by Fajardo et al. compared performance on un-cued items with their multiple-choice counterparts in summative evaluation in a radiology clerkship.⁵ They reported an adequate Cronbach's alpha coefficient of 0.77 for scores obtained on un-cued version of the items (Cronbach's alpha for multiple-choice version of the items was not reported) and discussed the importance of reliability in measurement precision and pass-fail decisions on high stakes assessments. The study by McManus et al. assessed reliability of the Member of Royal College of Physicians (MRCP) Part I postgraduate exam given in the United Kingdom.¹² They also studied how reliability is related to mean score and spread (standard deviation) of examinee scores. Average Cronbach's alpha coefficient of the exams was 0.86 (range: 0.83–0.89, standard deviation: 0.018).¹² Multiple regression analysis was conducted, which showed that

reliability tended to be higher when mean score on the exam or its standard deviation was high.¹² McManus et al. concluded that reliability is related to the mean and spread of exam scores.¹² Hutchinson et al. published a systematic review of literature on reliability of eleven postgraduate exams in the US, UK, Canada and Israel.¹³ Assessment instruments in these exams ranged from multiple-choice tests to oral tests. Reliability coefficients were found to range between 0.55 – 0.96, with a median coefficient of 0.77. Hutchinson et al. also discussed the importance of high internal consistency reliability in rendering meaningfulness to scores obtained on postgraduate exams.

The above studies emphasize the importance of enhancing the reliability of high-stakes exams. Reliability of scores on such exams can be improved by increasing the number of items in the exams.¹¹ Improvement expected from adding items can be estimated using the Spearman-Brown “prophecy” formula:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\text{sum of variances of all items}}{\text{total test variance}} \right),$$

where “ α ” is the Cronbach’s alpha coefficient and “ k ” is the number of items in an exam¹⁴. However, the number of items given in high stakes in-house or licensure exams is usually fixed and is usually not meddled with just for the sake of reliability. An alternate way to improve reliability of scores, apparent from the Spearman-Brown prophecy formula, is to increase total test variance by spreading out the subject performance. This can be achieved by increasing the difficulty of exam items, as discussed by McManus et al.¹² Increasing the difficulty of exam items helps in eliciting a wider range of performances among examinees thereby increasing the standard deviation, hence variance of observed scores.¹¹ An increased value of the denominator (total test variance) in the Spearman-Brown prophecy formula shown above would raise the value of Cronbach’s alpha coefficient.¹⁴ The positive effect of increased

standard deviation was also dwelled upon by McManus et al., who noted that reliability of high stakes assessment tends to increase whenever greater spread (standard deviation) of examinee scores is noted.¹²

How distractor functioning in multiple-choice questions impacts the spread and reliability of scores is also discussed in this chapter of the dissertation. The purpose of dwelling on this connection is to advance our understanding of the role of quality of multiple-choice questions in producing reliable scores on high-stakes exams. Research hypothesis of this part of the study was: Increased distractor functioning improves reliability of scores obtained on multiple-choice questions.

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 experiment group comprised students receiving the multiple-choice version, while those receiving the free-response version served as controls. 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. Table II-2 displays gender representation (percentage of male and female students), grade point average (GPA) in undergraduate studies, and average medical college admissions test (MCAT) scores in each cohort. Some degree of variation in all these characteristics was seen across the

Table II-2. Demographic characteristics of each cohort of subjects.

	Gender representation (% of males, % of females)	Average undergraduate (pre-matriculation) GPA	Average Medical College Admissions Test score
Cohort 1	53.2%, 46.8%	3.62	27.6
Cohort 2	36.4%, 63.6%	3.72	27.2
Cohort 3	56.1%, 43.9%	3.65	28.4
Cohort 4	52.3%, 47.7%	3.67	28.5
Cohort 5	58.4%, 41.6%	3.71	28.0
Cohort 6	52.9%, 47.1%	3.71	27.8

cohorts, which may represent fluctuation in trends and institutional policies in regards to medical school admissions.

The school's undergraduate medical education curriculum is a hybrid of Patient-Centered Learning (PCL) as well as traditional, discipline-based instruction. In this institution, neurohistology is taught during the neuroscience curricular block, which is scheduled at the end of academic Year 1. Instruction in neurohistology takes place through lectures and laboratory exercises conducted 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 in this study. Two versions the exam were used: the free-response (fill-in-the-blank format) version and the multiple-choice (one-best format with one correct option and 2, 3 or 4 incorrect options) version. The only difference between free-response and multiple-choice versions laid in the format of the asked question (example: Figure II-1). Of the 25 free-response – multiple-choice item sets, two item-sets were excluded from analysis since their free-response version contained options, thereby not meeting the criterion needed for comparison with the multiple-choice version.

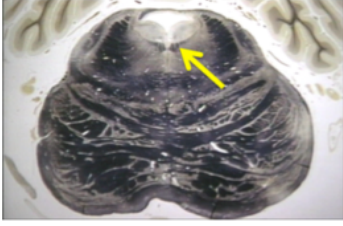
<p>Free-response (FR) version of sample item:</p> <p>Name the clinical finding that would be associated with an infarct in the indicated structure.</p>		<p>Multiple-choice (MCQ) version of sample item:</p> <p>A lesion of the indicated structure will result in which of the following clinical findings?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

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

Procedure

Each cohort of students was invited, via email, to attend a non-mandatory practice session for the end of curricular block Neurohistology exam scheduled approximately five days later. No information in regards to design of the study was shared in advance. No consent was sought and no points were granted for participation in the study. Once seated, free-response and multiple-choice versions of the exam printouts were randomly distributed. Then, the purpose of the study was shared, and students were asked not to provide any personal or identifiable information on the answer sheets. Neurohistology images were projected on a big 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

Based on performance in Cohorts 1 – 4, the multiple-choice version of the exam underwent the following two revisions.

- a. Thirty-one distractors in 15 multiple-choice questions with previous selection frequency of 0% were replaced with new distractors developed from frequent incorrect responses on free-response version of the items.

- b. Five 5-option multiple-choice questions were converted to 4-option multiple-choice questions via removal of a distractor with 0% selection frequency. The number of 5-, 4- and 3-option multiple-choice questions in the original (unrevised) version was 21, 1 and 1, respectively. The number of 5-, 4- and 3-option multiple-choice questions in the revised version was 16, 6 and 1 respectively.

The revised multiple-choice version of the exam was given to all subjects in Cohort 5. The purpose thereof was to note the extent of distractor functioning in the revised multiple-choice version of the exam from a bigger sample of subjects. In the next cohort (Cohort 6), the revised version was given to a random half of subjects (as in Cohorts 1 – 4) for the purpose of assessing the difference between *expected* and *observed* MCQ difficulty indices; the other half of Cohort 6 received the free response version of the exam.

Data Collection and Analysis

The following variables were calculated from student performance.

- a. Individual scores, as well as each cohort's mean and standard deviation.
- b. Psychometric characteristics of each item, i.e., item difficulty index (free-response, *expected* multiple-choice and *observed* multiple-choice) and point biserial correlation. Difficulty index is defined as the proportion of examinees answering the item correctly and is calculated as follows: number of correct answers / number of all answers.¹⁵ Point biserial correlation (a.k.a. item-total correlation) is the correlation coefficient of scores on an individual item with the sum of scores obtained on all other items.¹⁵ Point biserial correlation

ranges from – 1.00 to + 1.00; a higher positive value means that performance on an item correlates well with overall performance on an exam and indicates greater discriminatory ability of the item.¹⁵ On the other hand, a value of zero indicates that performance on an item has no correlation with overall performance on an exam, while a higher negative value means that performance on an item correlates inversely with overall performance on an exam.¹⁵

- c. According to accepted definitions, when an item has low point biserial correlation (usually < 0.2), it is considered to be a poor measurement of the intended construct and such an item is flagged for revision or removal.¹⁵
- d. Effect size of the differences between *expected* and *observed* MCQ difficulty indices. Effect size is an index of 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.¹⁶ Effect size was calculated via Cohen's *d*. 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 $d = \frac{M_e - M_c}{\sigma_c}$).¹⁶
- e. Total number of distractors with $\geq 5\%$, $\geq 10\%$, $\geq 20\%$, and $\geq 33\%$ selection frequency in the multiple-choice version of the exam in each cohort. Also noted were the numbers of total and functioning ($\geq 5\%$ selection frequency) distractors per multiple-choice question.

- f. Cronbach's alpha of scores obtained on multiple-choice version of the exam, before and after revision. It is conventionally accepted among psychometric scholars that a coefficient of at least 0.8 is satisfactory for high-stakes exams.^{1, 11, 14}
- g. Standard Error of Measurement ($SEM = SD\sqrt{1 - reliability}$), which is the standard deviation of an examinee's observed score given her/his true score.¹² As discussed in the introduction, "true" score of an examinee is the score that is uninfluenced by error.¹² Standard Error of Measurement describes precision of measurement and is used to establish a confidence interval within which an examinee's true score is expected to fall.¹² Note: 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.¹⁷

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

Results

Difficulty indices and effect size of the difference between *expected* and *observed* multiple-choice difficulty indices are discussed in the sub-section of Results titled "Validity of scores obtained on multiple-choice questions". Cronbach's alpha coefficient is discussed in the sub-section of Results titled "Reliability of scores obtained on multiple-choice questions".

Validity of Scores Obtained on Multiple-choice Questions

Table II-3 displays the number of students taking the free-response and multiple-choice versions of the exam, score means and their standard deviations, mean item difficulty indices and mean point biserial correlations. As expected, scores on free-response version of the exam tended to be lower in all cohorts than scores on multiple-choice version of the exam.

Table II-3. 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 shown.

	Cohort 1		Cohort 2		Cohort 3		Cohort 4		Cohort 5	Cohort 6	
	FR	MC Q	FR	MC Q	FR	MC Q	FR	MC Q	MCQ (only)	FR	MC Q
Number of students taking the exam	28	31	27	31	30	23	28	27	71	34	33
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

On the free-response version, mean item difficulty indices ranged from 0.64 to 0.70 and the mean point biserial correlation ranged from 0.30 to 0.42. On the multiple-choice version of the exam, before replacement of previously non-functioning distractors (Cohorts 1 – 4), mean item difficulty indices ranged from 0.82 to 0.85 and the mean point biserial correlation ranged from 0.25 to 0.43. Difficulty on the revised multiple-choice version (after replacement of previously non-functioning distractors) was found to be higher (lower difficulty index value) than seen in earlier cohorts. In Cohort 5, in which all students took the revised multiple-choice version, mean difficulty index was noted to be 0.74. In Cohort 6, in which a random half of the examinees took the revised multiple-choice version, it was noted to be 0.79. Mean point biserial correlation on the revised multiple-choice version was found to be 0.38 in Cohort 5 and 0.39 in Cohort 6; these

values were higher compared to those seen in the previous three cohorts (range: 0.25 – 0.30). These results show that, overall, the revised multiple-choice version was more difficult and offered slightly better discrimination amongst students of different abilities than the unrevised multiple-choice version of the exam.

Table II-4 displays mean *expected* and *observed* multiple-choice difficulty indices before (Cohorts 1 – 4) and after (Cohort 6) replacement of previously non-functioning distractors. The index of *expected* MCQ difficulty could not be calculated for Cohort 5, since all students in that cohort received the revised multiple-choice version of the exam, and no free-response difficulty index was available to calculate the index of *expected* multiple-choice difficulty. Effect size (Cohen’s *d*) of the difference between mean *observed* and *expected* multiple-choice difficulty indices, as well as the number of total and average functioning distractors (per MCQ) are also displayed in Table II-4.

Table II-4. Mean Free-Response (FR) and Multiple-Choice (MCQ) difficulty indices 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 difficulty index (Standard Deviation)	0.7 (0.20)	0.67 (0.25)	0.64 (0.23)	0.69 (0.18)		0.68 (0.22)
Mean <i>Expected</i> MCQ difficulty index (Standard Deviation)	0.78 (0.15)	0.76 (0.19)	0.74 (0.17)	0.77 (0.14)		0.76 (0.16)
Mean <i>Observed</i> MCQ difficulty index (Standard Deviation)	0.85 (0.13)	0.83 (0.17)	0.82 (0.19)	0.85 (0.14)	0.74 (0.16)	0.79 (0.16)
Effect size (Cohen’s <i>d</i>) of the difference b/w Mean <i>Observed</i> and <i>Expected</i> MCQ difficulty indices	0.46	0.40	0.46	0.59		0.15

Before replacement of previously non-functioning distractors (Cohorts 1 – 4), mean *expected* MCQ difficulty index ranged from 0.74 to 0.78 with a standard deviation ranging from 0.14 to 0.19. After replacement of previously non-functioning distractors (Cohort 6), mean *expected* MCQ difficulty index was noted to be 0.76, with a standard

deviation of 0.16. It is worth remembering that the index of *expected* MCQ difficulty was calculated using the difficulty index of free-response version of the items. The finding that the mean index of *expected* MCQ difficulty and its standard deviation tended to be similar before and after replacement of previously non-functioning distractors shows that difficulty of free-response version of the exam did not experience much change across the six cohorts. The mean *observed* MCQ difficulty index before removal of the non-functioning distractors (Cohorts 1 – 4) ranged from 0.82 to 0.85 with standard deviations ranging from 0.13 to 0.19. After replacement of previously non-functioning distractors, mean *observed* MCQ difficulty was noted to be 0.74 (Cohort 5) and 0.76 (Cohort 6) with standard deviations of 0.16 in both cohorts. This shows that, overall, the revised multiple-choice version of the exam was more difficult than the unrevised version. Before replacement of previously non-functioning distractors (Cohorts 1 – 4), the disparity between mean *expected* and *observed* MCQ difficulty indices was noted to be 7 – 8%; effect size (Cohen's *d*) of this difference ranged between 0.40 – 0.59. After replacement of previously non-functioning distractors (Cohort 6), the disparity between mean *expected* and *observed* MCQ difficulty indices was noted to be 3%; effect size (Cohen's *d*) of this difference was noted to be 0.15. This shows that the disparity between mean *expected* and *observed* MCQ difficulty indices reduced considerably after replacement of previously non-functioning distractors. Figure II-2 illustrates the effect size of the difference between mean *expected* and *observed* MCQ difficulty indices in all cohorts.

Table II-5 displays the number of total distractors as well as the number of distractors with $\geq 5\%$, $\geq 10\%$, $\geq 20\%$, and $\geq 33\%$ selection frequency in the multiple-choice

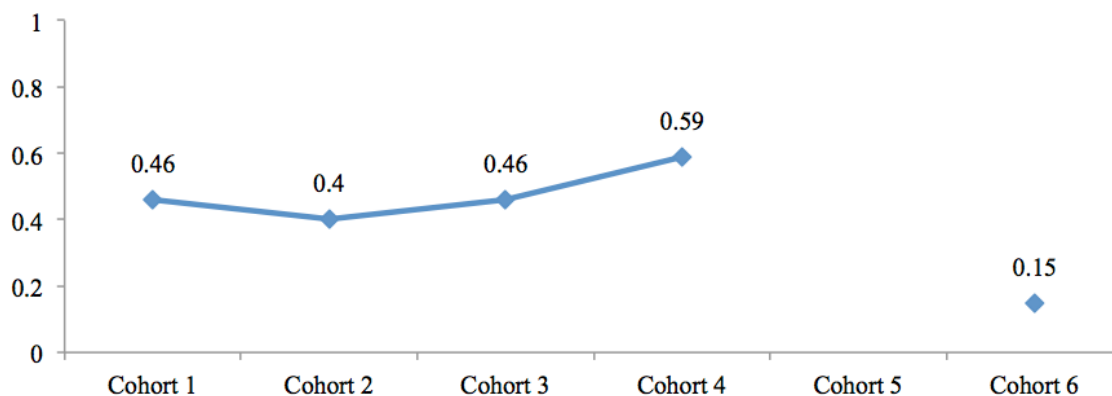


Figure II-2. Effect size (Cohen's d) of the difference between mean expected and observed MCQ difficulty indices. Effect size could not be calculated from Cohort 5 since all students received the revised MCQ version of the exam.

Table II-5. Number of total MCQ distractors as well as the number of distractors with $\geq 5\%$, $\geq 10\%$, $\geq 20\%$ and $\geq 33\%$ selection frequency in each cohort. Number of total and functioning ($\geq 5\%$) distractors per MCQ in each cohort is also displayed.

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

version of the exam, before (Cohorts 1 – 4) and after (Cohorts 5 and 6) replacement of previously non-functioning distractors. Before replacement of previously non-functioning distractors (Cohorts 1 – 4), the number of distractors with $\geq 5\%$ selection frequency ranged from 22.47% to 26.97%, while after replacement of previously non-functioning distractors, it was found to be 45.24% (Cohort 5) and 27.38% (Cohort 6). The number of distractors with $\geq 10\%$ selection frequency was to found range from 6.74% to 14.61% before (Cohorts 1 – 4), and 25% (Cohort 5) and 16.67% (Cohort 6) after replacement of previously non-functioning distractors. The number of distractors with

$\geq 20\%$ selection frequency was found to range from 2.25% to 6.74% before (Cohorts 1 – 4), and 5.95% (Cohort 5) and 8.83% (Cohort 6) after replacement of previously non-functioning distractors. Finally, the number of distractors with $\geq 33\%$ selection frequency was to found range from 0% to 5.62% before (Cohorts 1 – 4), and 0% (Cohort 5) and 3.57% (Cohort 6) after replacement of previously non-functioning distractors.

Table II-5 also displays the number of total and functioning ($\geq 5\%$) distractors per multiple-choice question. The number of distractors per MCQ was found to range from 0.87 to 1.04 before (Cohorts 1 – 4), and 1.65 (Cohort 5) and 1.00 (Cohort 6) after replacement of previously non-functioning distractors.

One finding becomes clear from the distractor selection frequency data shown in Table 6. The finding is that a revised multiple-choice version of the exam (Cohorts 5 and 6) displays higher distractor selection in most categories than unrevised multiple-choice version of the exam (Cohorts 1 – 4); the revision entailed replacement of previously non-functioning distractors with those developed from incorrect responses on free-response version of the items. This shows that the afore-mentioned revision of the multiple-choice version of the exam helped elicit greater distractor selection from the examinees. The effect of this greater distractor selection was increased mean multiple-choice question difficulty (Tables II-3 and II-4) and discriminatory ability (Table II-3), as well as reduction in the effect size of the difference between expected and observed multiple-choice difficulty indices (Table 5). Figure II-3 displays the percentage of multiple-choice distractors with different selection frequencies ($\geq 5\%$, $\geq 10\%$, $\geq 20\%$ and $\geq 33\%$).

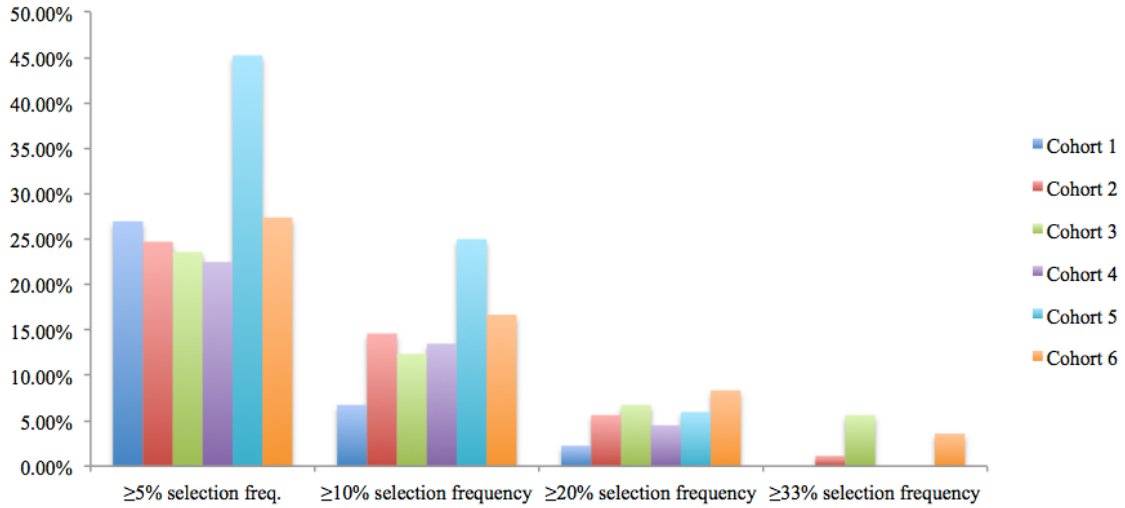


Figure II-3. Percentage of MCQ distractors with different selection frequencies. Among the cohorts that received both the FR and MCQ versions of the exam. Cohort 6 displayed higher distractor selection in most categories.

Reliability of Scores Obtained on Multiple-choice Questions

Table II-6 displays the number of students taking the free-response and multiple-choice versions of the exam, their mean scores, standard deviations, reliability coefficients (Cronbach's alpha) and Standard Errors of Measurement (SEM). Again, please note that all examinees in Cohort 5 received the revised multiple-choice version of the exam, and therefore performance data on free-response version of the exam is not available in this cohort. Before replacement of previously non-functioning distractors (Cohorts 1 – 4), the standard deviations of scores obtained on free-response version of the exam tended to be greater in most cohorts (Cohorts 2, 3 and 4) than the standard deviation of scores on the (unrevised) multiple-choice version of the exam. However, after replacement of previously non-functioning distractors (Cohort 6), the standard deviation of scores obtained on the (revised) multiple-choice version of the exam was found to be greater than that on the free-response version. This shows that, generally, the free-response version of the exam was able to elicit a greater range of abilities from

Table II-6. Number of students taking the FR and MCQ versions of the exam, mean score, standard deviation, reliability coefficient (Cronbach's alpha) and Standard Error of Measurement (SEM) 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 students	28	31	27	31	30	23	28	27	71	34	33
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
Range of scores	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

examinees than the unrevised multiple-choice version that had displayed lower distractor functioning (Cohorts 1 – 4, Table 6). However, a reversal in this pattern was observed when various non-functioning distractors were replaced. The range of ability elicited by the revised multiple-choice version, with greater distractor functioning (Cohort 6, Table II-6), was found to be greater than that elicited by the free-response version. The presence of greater standard deviation of scores, hence greater range of observed abilities amongst examinees, has implication on the reliability coefficient of scores as discussed in the following paragraphs.

Before replacement of previously non-functioning distractors (Cohorts 1 – 4), mean scores on the (unrevised) multiple-choice version of the exam were found to range between 18.80 and 19.60. After replacement of previously non-functioning distractors (Cohorts 5 and 6), mean scores on the (revised) multiple-choice version of the exam were found to be 17.04 (Cohort 5) and 18.24 (Cohort 6). Similarly, standard deviations on unrevised multiple-choice version of the exam were found to range between 2.11 and 3.34, while on the revised multiple-choice version of the exam, a higher standard deviation of 3.61 was noted (Cohorts 5 and 6).

Here is a narrative about Cronbach's alpha coefficient and Standard Error of Measurement values shown in Table II-6. Before replacement of previously non-functioning distractors (Cohorts 1 – 4), the reliability coefficient (Cronbach's alpha) on the multiple-choice version of the exam was noted to range between 0.43 and 0.80, while after replacement of non-functioning distractors, it was noted to be 0.74 (Cohorts 5) and 0.78 (Cohort 6). Values of Cronbach's alpha coefficient observed on the multiple-choice version of the exam after replacement of non-functioning distractors (Cohorts 5 and 6) were found to be higher than the ones seen on multiple-choice version of the exam in the previous three cohorts (Cohorts 2, 3 and 4) before distractor replacement. Also after replacement of non-functioning distractors, a slightly higher Standard Error of Measurement (1.84 in Cohort 5 and 1.66 in Cohort 6) on multiple-choice version of the exam was noted; before replacement of non-functioning distractors (Cohorts 1 – 4), the Standard Error of Measurement was noted to range between 1.49 and 1.59.

Figure II-4 demonstrates the relationship between standard deviation of scores and their reliability coefficient based on the data presented in Table II-6. Whenever scores obtained on the multiple-choice version of the exam exhibited greater standard deviation (Cohorts 1, 5 and 6), the reliability coefficient (Cronbach's alpha) was also noted to be higher. Greater standard deviation of scores on the multiple-choice version of the exam in Cohorts 5 and 6 is attributable to increased distractor functioning noted in these cohorts (Table II-5, Figure II-3). However, higher standard deviation of scores on the multiple-choice version of the exam seen in Cohort 1 is an interesting finding, since students in that cohort received the unrevised multiple-choice version of the exam that

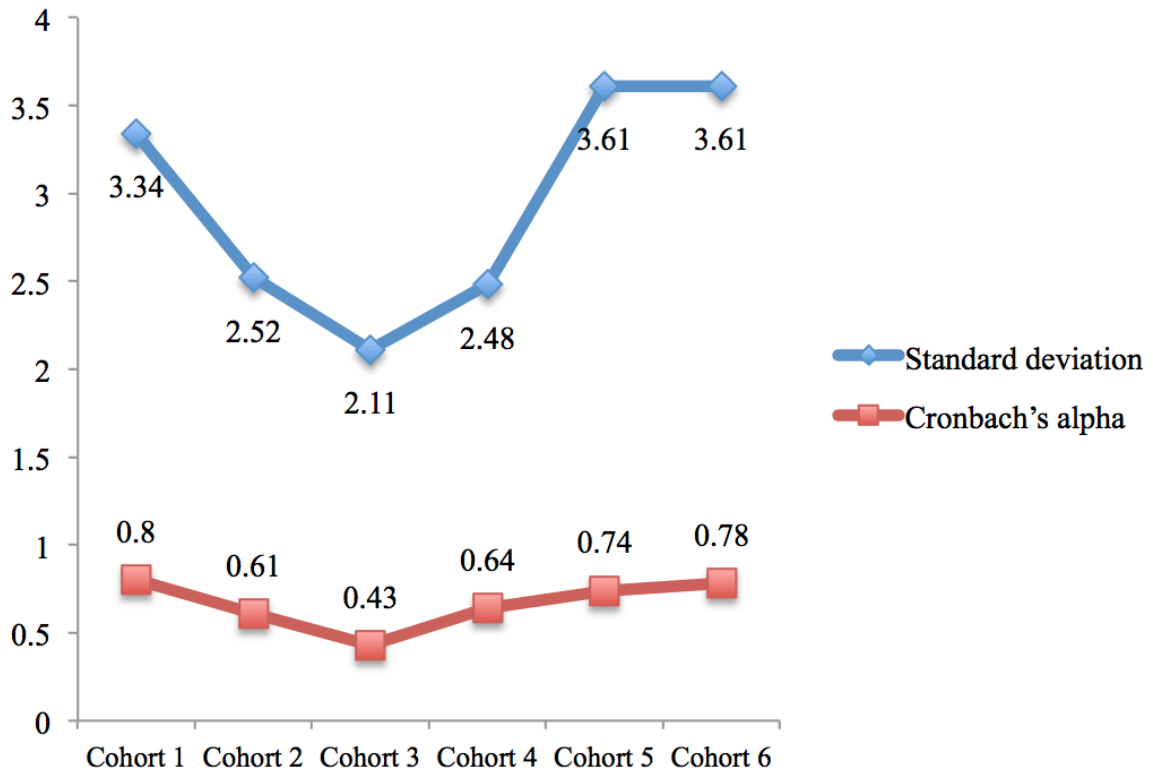


Figure II-4. Standard deviation and reliability coefficient (Cronbach's alpha) of scores obtained on multiple-choice version of the exam.

had displayed lower overall distractor functioning than the rest of cohorts. A possible explanation of this finding follows.

Table II-6 also presents data on the range of scores seen on the free-response as well as the multiple-choice version of the exam in all cohorts. The range of scores on the unrevised multiple-choice version of the exam seen in Cohort 1 was noted to be greater than the range seen in other cohorts (Cohorts 2 – 4) that also took the unrevised multiple-choice version. While the maximum score obtained on multiple-choice version of the exam was the same across all cohorts, there was a difference in the minimum score. A minimum score of 8 was observed in Cohort 1, while it was 12, 14 and 13 in Cohorts 2 – 4, respectively. This shows that, for some reason, there were some very low performing

student in Cohort 1 whose lower scores increased the range of ability (standard deviation) observed in Cohort 1. Looking at the mean item difficulty on the multiple-choice version of the exam in Cohort 1 (Table II-3), one can surmise that these low-performing students performed poorly on an otherwise easier exam, likely due to their lack of preparedness. This lack of preparedness negatively impacted these students' scores thereby increasing the standard deviation of exam scores. And, owing to the directly proportional relationship between the standard deviation and the reliability (Cronbach's alpha) coefficient ($\alpha = \frac{k}{k-1} (1 - \frac{\text{sum of variances of all items}}{\text{total test variance}})$),¹⁴ the reliability of scores on the unrevised multiple-choice version of the exam in Cohort 1 was noted to be higher than Cohorts 2, 3 and 4 that also took the unrevised version.

Another peculiar finding was a slight increase in the standard error of measurement (SEM) of scores on the multiple-choice version of the exam after replacement of non-functioning distractors (Cohorts 5 and 6) (Table II-6). As discussed in the Methods section of this chapter, the Standard Error of Measurement is the standard deviation of the observed score given an examinee's true score, and thus provides an estimate of measurement precision. An explanation for the increase in the Standard Error of Measurement after replacement of previously non-functioning distractors is the directly proportional relationship between the standard deviation of scores and the Standard Error of Measurement.^{12, 17} The equation describing this relationship in psychometrics literature is $SEM = SD\sqrt{1 - \text{reliability}}$, in which SD is the standard deviation of scores.^{12, 17} This equation shows that while an increased range of ability (standard deviation) elicited by an exam increases the reliability coefficient of obtained scores, it also increases the error of measurement when the reliability is held constant.

This directly proportional relationship is a likely explanation for higher the Standard Error of Measurement noted on revised multiple-choice version of the exam (Cohorts 5 and 6).

Discussion

The purpose of this study was to demonstrate the impact of distractor functioning on validity and reliability of scores obtained on multiple-choice version of an exam compared to its free-response version. Revision of the multiple-choice version of the exam, via replacement of consistently non-functioning distractors with those developed from incorrect responses on free-response version of the items, was carried out. An index of *expected* MCQ difficulty was calculated via the difficulty index of the free-response version of the item and the number of options provided in the multiple-choice version of the item (Table II-1). A component of validity of scores obtained on multiple-choice version of the exam was assessed via comparison of the index of *expected* MCQ difficulty with the index of *observed* MCQ difficulty before and after replacement of previously non-functioning distractors. Reliability of multiple-choice version of the exam was assessed via calculation of Cronbach's alpha coefficient of scores obtained before and after replacement of previously non-functioning distractors. A few observations can be made from the obtained results.

Firstly, performance on the free-response version of an exam of neurohistology knowledge was consistently lower than performance on its multiple-choice version (Table II-2). Since the free-response and multiple-choice versions were randomly distributed in each cohort, the consistently disparate performance can be attributed to the version of the exam. The multiple-choice version of the items contains options, which

allow some degree of cueing and correct guessing from examinees, thereby leading to higher mean scores. This means that an examinee can correctly answer a multiple-choice question by recognizing the correct option without producing the answer spontaneously from memory. The free-response (fill-in-the-blank) format of assessment greatly reduces the potential for such guessing and cueing. Discussion on the guessing and cueing phenomenon exclusive to the format of multiple-choice questioning can be found in the literature.¹⁸⁻²²

Secondly, the average difficulty index of items in a multiple-choice exam is below its expected value when the number of distractors with sufficient plausibility ($\geq 5\%$, $\geq 10\%$, $\geq 20\%$ and $\geq 33\%$ selection frequencies) is low. Tables II-4 and II-5 highlight this finding. Effect size of the difference between mean *expected* and *observed* multiple-choice difficulty indices was found to be higher in cohorts with lower overall distractor functioning (Cohorts 1 – 4). It is worth noting that most students rule-in or rule-out various distractors based on their partial knowledge of the content under assessment. Therefore, psychometric scholars advise inclusion of only those multiple-choice distractors that reasonably elicit such partial knowledge from the examinees.^{3, 8} The commonly used criterion for such reasonable elicitation of partial knowledge is minimum 5% selection frequency of a distractor.^{3, 8} In the study presented in this chapter of the dissertation, a dearth of such reasonable, justifiably included distractors was seen in the unrevised multiple-choice version of the exam (Cohorts 1 – 4) (Table II-5, Figure II-3). This, we believe, was the reason behind the disparity between *expected* and *observed* MCQ difficulty indices. In other words, item writers' ability to accurately gauge student knowledge was compromised by lack of plausible distractors, which

allowed the test wise (and not necessarily *well-prepared*) students to perform well beyond expectation. Low distractor functioning rendered the multiple-choice questions to be easier than expected, thereby reducing the validity evidence of obtained scores. However, when distractor functioning was increased via inclusion of more plausible distractors, such as those developed from incorrect responses on free-response version of the same items, the observation changes considerably. Table II-5 shows that after replacement of previously non-functioning distractors (Cohorts 5 and 6), an increase in the number of multiple-choice distractors with $\geq 5\%$, $\geq 10\%$, $\geq 20\%$ and $\geq 33\%$ selection frequencies is noted. The increased distractor functioning, in turn, reduces the disparity between average *expected* and *observed* multiple-choice difficulty indices (Table II-4, Cohort 6), thereby strengthening the evidence of validity of scores obtained on the multiple-choice exam. The above argument is strengthened by previously published reports that difficulty of multiple-choice items is contingent upon quality, not quantity, of its distractors.^{9, 20} Overall, our finding affirms the notion that inclusion of distractors with greater plausibility (hence, “functioning”) is vital for reducing the much-dreaded cueing effect and amelioration of quality of multiple-choice assessment. A study published by Prihoda et al. demonstrated improvement in validity of scores obtained on multiple-choice exams in a somewhat different fashion.⁶ Their intervention entailed post-hoc correction for guessing of multiple-choice scores, leading to greater agreement (intraclass correlation coefficient) with scores obtained on free-response version of the same exam. The study presented in this chapter of the dissertation, however, uses an active intervention (replacement of non-functioning distractors with more plausible, functioning ones) to generate the evidence of validity. This approach is yet to be reported

in literature in the context of assessment in undergraduate medical education, which highlights the novelty of the presented method of assessing the quality of multiple-choice exams in undergraduate medical education.

Thirdly, when distractors developed from incorrect responses on free-response version of the items are used to replace consistently non-functioning distractors, an increase in average discriminatory ability of multiple-choice questions is noted. Table II-3 highlights this finding; average difficulty index was found to range between 0.82 – 0.85 and point biserial correlations were found to range between 0.25 – 0.40 before revision of the multiple-choice version of the exam (Cohorts 1 – 4). After the revision (replacement of non-functioning distractors), the multiple-choice version showed average difficulty index of 0.74 (Cohort 5) and 0.79 (Cohort 6), and point biserial correlation of 0.38 (Cohort 5) and 0.39 (Cohort 6). This increase in difficulty and discriminatory ability of the multiple-choice version of the exam occurred in the setting of increased functioning, i.e. selection, of its distractors (Table II-5 and Figure II-3). This affirms the notion that plausible distractors gauge conceptual misunderstandings more accurately, allowing better separation of low- and high-ability students.

Fourthly, increased distractor functioning enhances the reliability of scores obtained on multiple-choice questions (Table II-6). After replacement of non-functioning distractors (Cohorts 5 and 6), the revised multiple-choice version of the exam exhibited greater distractor functioning resulting in a lower mean, greater range, and higher standard deviation of scores. Consequently, an increase in the reliability coefficient (Cronbach's alpha) was noted owing to the directly proportional relationship between standard deviation and the reliability coefficient

$(\alpha = \frac{k}{k-1} (1 - \frac{\text{sum of variances of all items}}{\text{total test variance}}))$.¹⁴ As apparent from the quoted formula, reliability of scores increases when the range of ability (spread of scores or “test variance”) elicited by an exam is increased, other things being equal. The revised multiple-choice version of the exam (Cohorts 5 and 6) helped in eliciting a greater range of abilities from examinees via provision of higher quality multiple-choice questions with enhanced distractor functioning. Resultantly, the reliability of scores obtained on the revised multiple-choice version of the exam was noted to be higher than that seen in previous three cohorts that took the unrevised multiple-choice version of the exam.

It is worth noting, however, that reliability may be affected by factors other than the quality of multiple-choice questions, vis-à-vis distractor functioning. If, for some reason, more low-performing (low ability, or unprepared) students are encouraged to take an exam, their outlying (lower) performance may increase standard deviation, and consequently the reliability coefficient of obtained scores. And this may occur despite low quality of the overall exam, as evident from weak psychometric characteristics and low multiple-choice distractor functioning. This phenomenon, observed in Cohort 1 (Table II-6), has been observed and reported on by other scholars as well, who have argued that the judgment on reliability of an exam’s scores should not be based purely on the reliability coefficient. In a study published by Tighe et al., the interrelationships among standard deviation, Standard Error of Measurement and exam reliability of scores were investigated via a Monte Carlo simulation of 10,000 candidates taking a postgraduate exam.²² It was found that the very same exam dropped its reliability dramatically when retaken by only those examinees who had already passed it. This shows that when ability range of examinees is artificially restricted (such as when only

high ability examinees are allowed to retake an exam), the reliability coefficient of an exam tends to decrease. Conversely, reliability coefficient can become artificially inflated when very weak candidates are encouraged to take the exam. From the findings reported by Tighe et al., as well as those seen from performance on the (unrevised) multiple-choice version of the exam in Cohort 1, it becomes clear that there are instances when reliability coefficient is not a relevant measure of exam quality. Tighe et al. suggest that in cases where ability range of examinees is noted to be narrow, the Standard Error of Measurement may be enough for assessment of precision measurement. We agree with this suggestion and advise interpretation of the reliability coefficient in light of the psychometric characteristics (difficulty index and point biserial correlations) and degree of distractor functioning noted in a multiple-choice exam.

A few limitations to the findings exist. The small number of investigated items (n=23), although suitable for assessment of knowledge of histology, may be insufficient for an experiment of this nature. This issue was mitigated to some degree by similar findings noted in several cohorts (Cohorts 1 – 4) that took the unrevised multiple-choice version of the exam. The findings noted after revision (replacement of non-functioning distractors) of the multiple-choice version will need to be evaluated for consistency in future cohorts of subjects. Generalizability of our results to exams given in other basic science subjects, and to students undertaking other health science curricula, is yet to be evaluated; we invite scholars with diverse backgrounds to conduct similar studies in their domain. Another potential limitation is the no-stakes nature of the exam used in this study; it was given as practice for the high stakes neuroscience exam scheduled five days later for every cohort of subjects. Besides any influence of no-stakes nature of the exam

used in this study, it is worth noting that our research question focused solely on differential performance on free-response and multiple-choice versions of an exam at a single time point.

Multiple-choice questions are the norm in assessment in undergraduate medical education for their ease of administration and objective grading. However, it is clear from the results presented in this study that in-house multiple-choice assessment may rely on plain recognition of the most credible answer from a brief list of options, some of which are barely plausible. This is far cry from real-life situations medical professionals face every day. Although signs and symptoms of an illness allow for some cueing and educated guessing, patients do not present with five options for a healthcare provider to mull over.²³ 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 items, is a useful step in this direction and generates evidence of validity of scores obtained on multiple-choice exams. 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.²⁴⁻²⁷ In the authors' experience, administering two versions (free-response and multiple-choice) of the same exam as practice for a subsequent high stakes exam requires effort that yields positive feedback from learners. It allows learners to detect areas of needed improvement and for instructors to encourage deep, rather than superficial, learning strategies.²⁸ Moreover, such efforts yield valuable data to enhance one's educator scholarship.

For valid competency-based assessment, scores obtained on any assessment instrument must have evidence of sound internal structure (moderate difficulty, and sufficient discriminatory ability and reliability) and must be able to gauge the problem solving ability of students.^{23,27} In the wake of reports that multiple-choice questions may not be the only viable choice in this regard, the idea of other assessment instruments, such as the Extended-Matching Questions (EMQs), has been floated.²⁷ While much is going on in the realm of standardized testing conducted by bodies such as National Board of Medical Examiners, in-house assessment remains reliant on multiple-choice questions. Therefore, it is important to raise the awareness among basic medical science faculty of the steps that can improve the quality of multiple-choice questions in in-house exams. One such step focuses on distractor plausibility. From the study presented in this chapter of the dissertation, one can see that the goal is indeed achievable. When a committed effort is made, quality of multiple-choice assessment improves, with positive impact on an exam's ability to accurately and reliably serve its purpose.

Since assessment drives learning,²⁹ and quality assessment is one of the tenants of competency-based education,³⁰ it is hoped that investigations such this one will help in the identification of truly competent students and will help inform a sound remediation process for the rest. Contemplation of how best to assess the knowledge that matters for future physicians has ultimate benefit for medical profession as well as the society as a whole.

References

1. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am J Med.* 2006;119(166):167–116.
2. Kern DE, Thomas PA, Hughes MT. Curriculum development for medical education: A six-step approach, 2nd edit. Baltimore, MD: Johns Hopkins University Press; 2009.
3. Haladyna TM, and Downing SM. How many options is enough for a multiple-choice test item? *Educ Measure Issues Prac.* 1993;53:999–1009.
4. Damjanov I, Fenderson BA, Veloski JJ, Rubin E. Testing of medical students with open-ended, uncued questions. *Hum Pathol.* 1995;26:362–365.
5. Fajardo LL, Chan, KM. Evaluation of medical students in radiology: written testing using uncued multiple choice questions. *Invest Radiol.* 1993;28:964–968.
6. Prihoda TJ, Pinckard RN, McMahan CA, Jones AC. Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *J.Dent Educ.* 2006;70(4):378–386.
7. Newble DI, Baxter A, Elmslie RG. A comparison of multiple choice and free response tests in examinations of clinical competence. *Med Educ.* 1979;13:263–268.
8. Rodriguez, M.C. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educ Measure Issues Prac.* 2005;24(2):3–13.
9. Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurs Educ Today.* 2010;30(6):539–543.

10. De Champlain A. A primer on classical test theory and item response theory for assessment in medical education. *Med Educ.* 2010;44:109–117.
11. Downing SM. Reliability: On the reproducibility of assessment data. *Med Educ.* 2004;38:1006–12.
12. McManus IC, Mooney-Somers J, Dacre JE, Vale JA. Reliability of the MRCP (UK) Part I examination, 1984–2001. *Med Educ.* 2003;37:609–11.
13. Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Med Educ.* 2002;36:73–91.
14. Karras DJ. Statistical methodology: II. Reliability and validity assessment in study design, part A. *Acad Emerg Med.* 1997;4:64–71.
15. Tavakol M, Dennick R. Post-examination analysis of objective tests. *Med Teach.* 2011;33:447–458.
16. Hojat M., Xu G. A visitor's guide to effect sizes—statistical significance versus practical (clinical) importance of research findings. *Adv Health Sci Educ.* 2004;9(3):241–249.
17. Harvill LM. NCME Instructional module: standard error of measurement. *Educ Measure Issues Prac.* 1991;10(2):33–41.
18. Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Appl Psych Meas.* 1982;6(1):1–11.
19. Norman GR, Smith EKM, Powles AC, Rooney PJ, Henry NL, Dodd PE. Factors underlying performance on written tests of knowledge. *Med Educ.* 1987;21:297–304.
20. Schuwirth LWT, van der Vleuten CPM, Donkers HHL. A closer look at cueing effects in multiple-choice questions. *Med Educ* 1996;30:44–49.

21. Norman GR. Problem solving skills, solving problems and problem-based learning. *Med Educ.* 1988;22:270–86.
22. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The standard error of measurement is a more appropriate measure of quality in postgraduate medical assessments than is reliability: An analysis of MRCP(UK) written examinations. *BMC Med.* 2010;10:40.
23. Veloski JJ, Rabinowitz HK, Robeson MR, Young PR. Patients don't present with five choices: an alternative to multiple-choice tests in assessing physicians' competence. *Acad Med.* 1999;74:539–546.
24. Case SM, Swanson DB, Ripkey DR. Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. *Acad Med.* 1994;69(10 suppl):S1–S3.
25. Swanson DB, Case SM. Variation in item difficulty and discrimination by item format on Part I (basic sciences) and Part II (clinical sciences) of U.S. licensing examinations. In: Rothman A, Cohen R, editors. *Proceedings of the Sixth Ottawa Conference on Medical Education.* Toronto, Canada: University of Toronto Bookstore Custom Publishing;1995. p. 285–287.
26. Swanson DB, Holtzman KZ, Clauser BE, Sawhill AJ. Psychometric characteristics and response times for one-best-answer questions in relation to number and source of options. *Acad Med.* 2005;80(10 suppl):S93–S96;16.
27. Swanson DB, Holtzman KZ, Allbee K, Clauser BE. 2006. Psychometric characteristics and response times for content-parallel extended- matching and one-best-answer questions in relation to the number of options. *Acad Med* 81(10 suppl):S52–S55.

28. Entwistle N. Promoting deep learning through teaching and assessment: Conceptual frameworks and educational contexts. In:TLRP First Annual Conference. Teaching and Learning Research Programme. Leicester, UK; 9–10 November 2000. TLRP, Institute of Education, University of London, London, UK. Available from: <http://www.tlrp.org/pub/acadpub/Entwistle2000.pdf> [accessed 24 May 2014].
29. Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: An unavoidable truth? *Anat Sci Educ.* 2009;2:199–204.
30. Van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ.* 2005;39:309–317.

CHAPTER III

**THE IMPACT OF ITEM FLAWS, TESTING OF LOW COGNITIVE LEVEL,
AND LOW DISTRACTOR FUNCTIONING ON MULTIPLE-CHOICE
QUESTION QUALITY**

Abstract

Background

Item writing flaws (IWFs), testing of low cognitive level (CL) and non-functioning distractors (NFDs) hinder accurate assessment of student knowledge via multiple-choice exams. This study evaluated the impact of addressing these issues on the psychometric characteristics (difficulty and discriminatory ability) of multiple-choice questions (MCQs) used in high-stakes assessment in Year 1 medical education.

Method

A repeated-measures experimental design was used in the setting of four end-of-curricular-block exams in a Year 1 undergraduate medical curriculum. Fifty-five MCQs with too high difficulty (difficulty index <0.4), too low difficulty (difficulty index >0.8), or insufficient discriminatory ability (point biserial correlation <0.2) on previous administration were identified. These items were blindly placed in either the experimental or the control group. There were two experimental sub-groups. The items in Experimental Subgroup A underwent removal of item-writing flaws along with enhancement of tested cognitive level (21 MCQs), while the items in Experimental Subgroup B underwent replacement or removal of non-functioning distractors (11 MCQs). The control group of items (Group C) did not undergo any intervention

(23 MCQs). Item writing guidelines from National Board of Medical Examiners and published literature were utilized in the interventions.

Result

Post-intervention, the number of functioning distractors ($\geq 5\%$ selection frequency) was noted to increase from 0.67 to 0.81 per MCQ in Subgroup A, and from 0.91 to 1.09 per MCQ in Subgroup B. The number of MCQs with sufficient point biserial correlation was also noted to increase from 0 to 10 in Subgroup A, and from 0 to 6 in Subgroup B; the increase in both subgroups was found to be statistically significant. No significant change in difficulty indices was noted post-intervention in the experimental sub-groups. Upon re-administration, the psychometric characteristics of MCQs in the control group (Group C) did not experience any significant change.

Conclusion

Correction of item writing flaws, removal or replacement of non-functioning distractors, and enhancement of tested cognitive level positively impacts discriminatory ability of multiple-choice questions. This helps prevent construct-irrelevant variance from affecting the evidence of validity of scores obtained on multiple-choice questions.

Introduction

Assessment in undergraduate medical education, as well as education in other health professions, is heavily reliant on multiple-choice questions (MCQs) written by faculty members responsible for taught content. Quality of such in-house assessment has been reported as threatened because of lack of adequate faculty development in multiple-choice question construction.¹⁻³ A brief synopsis of a few relevant reports follows.

A study by Jozefowicz et al. investigated the quality of in-house exams used in three U.S. medical schools.³ In their study, multiple-choice questions used in pre-clinical medical education were rated on a five-point scale in which a rating of “1” meant testing of recall only and presence of technical flaws in the question, and “5” meant usage of clinical or laboratory vignette, requiring reasoning to answer and lack of technical flaws in the question. Third party raters made independent assessments of the quality of various multiple-choice questions, and a mean score of the quality of each multiple-choice question was calculated. Mean scores of questions written by the item-writers trained by National Board of Medical Examiners (NBME) were compared with the mean scores of questions written by items writers without NBME training. Analysis of a total of 555 questions found mean rating for all questions to be 2.39. The 92 questions written by NBME-trained item writers had a mean rating of 4.24, and the 463 questions written by faculty without formal NBME training had a mean score of 2.03. Significant difference ($p < 0.01$) between these two ratings was found hinting at the low quality of exams written by faculty without any formal training in item writing. It was suggested that the quality of such examinations could be significantly improved by providing formal item-writing training aimed at basic science faculty.

Similarly, a study by Masters et al. assessed multiple-choice questions used in test-banks accompanying selected nursing textbooks². A random sample of 2,913 questions was selected from 17 test banks. Questions were evaluated on the basis of adherence to guidelines for writing multiple-choice questions and also on the basis of cognitive level as defined by Bloom's taxonomy, as well as on the basis of the distribution of correct answers as A, B, C, or D. A total of 2,233 violations of item-

writing guidelines were found in those questions. Most of those violations were minor, but some were serious. A large number of questions (47.3%) were written at low cognitive level of plain factual recall, and a meager 6.5% were written at the higher cognitive level of “analysis”. No significant difference in the distribution of answers (among choices A – D) was found. Masters et al. suggested that in-house faculty must evaluate multiple-choice questions critically before using them in exams.

Studies by Jozefowicz and Masters et al. highlight the commonly perceived threats to quality of multiple-choice questions. These threats include item writing flaws, testing of lower cognitive function and non-functioning distractors. Item-writing flaws (IWFs) shown in Table III-1 are violations of published, commonly accepted, item writing guidelines meant to prevent test wiseness and irrelevant difficulty from influencing examinee performance on multiple-choice exams.⁴ Examples of the high prevalence of item flaws in health science exams and their impact on assessment of learning outcomes have been discussed in the literature.^{2, 3, 5} Downing, in a study published in 2005, studied the adverse consequences of flawed multiple-choice questions and reported that 10–15% of students who failed in-house exams would have passed if questions with flaws were removed from the examinations.⁵ Tarrant and Ware, in a study published in 2008, evaluated assessment practices in nursing education.⁶ They reported the proportion of flawed multiple-choice questions in high-stakes exams to be in the range of 28–75%. In their study, fewer examinees were found to pass the exams when flawed multiple-choice questions were removed during the post-hoc analysis. Moreover, a greater number of examinees were found to score $\geq 80\%$ marks on unflawed multiple-

Table III-1. List of IWFs, published by NBME3, with corresponding numerical codes used in this study.

Code	<i>Issues Related to Testwiseness</i>
1	Grammatical cues - one or more distractors don't follow grammatically from the stem
2	Logical cues - a subset of the options is collectively exhaustive
3	Absolute terms - terms such as "always" or "never" are in some options
4	Long correct answer - correct answer is longer, more specific, or more complete than other options
5	Word repeats - a word or phrase is included in the stem and in the correct answer
6	Convergence strategy - the correct answer includes the most elements in common with the other options
	<i>Issues Related to Irrelevant Difficulty</i>
7	Options are long, complicated, or double
8	Numeric data are not stated consistently
9	Terms in the options are vague (e.g, "rarely," "usually")
10	Language in the options is not parallel
11	Options are in a nonlogical order
12	"None of the above" is used as an option
13	Stems are tricky or unnecessarily complicated
14	The answer to an item is "hinged" to the answer of a related item

choice questions than on flawed and unflawed multiple-choice questions combined. The findings reported by Downing, and those reported by Tarrant and Ware show that item flaws can surreptitiously increase both the pass as well as the failure rate in high stakes exams.

Another issue discussed in the literature in the context of, but not classified as, item flaws is testing of lower (factual recall) rather than higher (application of knowledge) cognitive function.⁷ The main reason for emphasizing testing of higher cognitive level is that application of basic medical sciences to clinical situations requires higher order thinking and deductive reasoning beyond pure regurgitation of facts.⁷ Tarrant et al., in a study published in 2006, reported on the use of nearly 3,000 multiple-choice questions in nursing education assessment over a five-year period.⁸ They reported that questions testing lower cognitive function were significantly more likely to contain item-writing flaws than those testing higher cognitive levels. Newble,⁹ Maguire et al.,¹⁰ and Elstein¹¹ have also reported on the adverse effect of testing lower cognitive function,

and found it to hinder valid assessment of students' problem-solving ability.⁹⁻¹¹ The study by Newble et al. compared the performance of undergraduate medical students (novices) and practicing physicians (experts) on identical and equivalent tests written in multiple-choice and free-response formats.⁹ The tests were designed to assess clinical competence at the hospital intern level. Performance on multiple-choice version of the exams was found to be better than that on the free-response version in both groups of examinees. However, the difference in performance noted on free-response and multiple-choice versions of the exams among medical students was greater than that seen among practicing physicians. Newble et al. attributed the difference in performance among medical students to greater reliance on guessing and cueing offered by the multiple-choice questions. A questionnaire survey given in this study showed that students were aware of the deficiencies in multiple-choice testing and, a large majority believed that free-response testing gave a more accurate assessment of their clinical ability. Newble et al. concluded that in tests aimed at measuring higher cognitive thinking, such as clinical application of knowledge, multiple-choice questions appear to overestimate a candidate's ability to an extent that made them less suitable than free-response questions for assessment of clinical competence. Others scholars have discussed ways to address the concern raised by Newble et al., including ways to effectively assess higher order thinking, clinical reasoning and problem-solving ability via carefully constructed multiple-choice questions.^{10, 12, 13}

The third topic of interest in the study presented in this chapter of the dissertation is distractor functioning. A *functioning* distractor (FD) is an incorrect option that is selected by $\geq 5\%$ of examinees (i.e., $\geq 5\%$ selection frequency).¹⁴ Another property

desirable in a functioning distractor is that it should be chosen more by low-performing examinees than high-performing examinees.¹⁴ Such selective attractiveness for low-performing students renders “negative” discriminatory ability to that distractor, which is a desired trait in a functioning distractor.¹⁴ On the other hand, a *non-functioning* distractor (NFD) is an incorrect option chosen by fewer than 5% of examinees and possesses a positive discriminatory ability, both of which are undesirable characteristics in a multiple-choice distractor.¹⁴ Low distractor functioning has been reported to threaten the validity of scores obtained on multiple-choice questions.^{14, 15} Tarrant et al. have reported on the impact of eliminating a non-functioning distractor from a 4- or 5-option multiple-choice question.¹⁶ The aim of their study was to study the effect of such removal on psychometric properties (difficulty and discriminatory ability) of multiple-choice questions. Using item-analysis data, they eliminated the distractor with the lowest selection frequency and compared the performance on 3- and 4-option versions of 41 multiple-choice questions in two cohorts of nursing students. They found that removing the non-functioning distractor resulted in minimal changes in item difficulty and discriminatory ability. The three-option version of the items were found to contain more functioning distractors despite having fewer distractors overall. Moreover, existing distractors were found to be more discriminatory when infrequently selected distractors were removed from the questions. Since three-option questions require less time to develop and administer, Tarrant et al. encouraged adoption of three-option multiple-choice questions as the standard in multiple-choice testing. Similarly, in a seminal meta-analysis published in 2005, Rodriguez utilized item-analysis data to eliminate the least functioning distractor from 41 four-option multiple-choice questions.¹⁵ He found no

significant difference in difficulty of multiple-choice question after removal of the least functioning distractor. He also reported that 4-option version of various multiple-choice questions contained less functioning distractors and possessed lower discriminatory ability than their 3-option versions. The conclusion was that eliminating a non-functioning distractor from a 4- or 5-option multiple-choice question does not impact its difficulty and discriminatory ability significantly and lends the benefit of reduced response time and increased content sampling for an exam.

The study presented in this chapter of the dissertation evaluates the impact of addressing item flaws, testing of low cognitive function and non-functioning distractors on quality of multiple-choice questions. The research question of the study was, “What is the effect of correction of item writing flaws (including testing at a higher cognitive level) and removal or replacement of non-functioning distractors on difficulty and discriminatory ability of multiple-choice questions?” The conceptual framework used in this study was validity, as defined by Messick¹⁷ and advanced by others,^{18, 19} in which evidence from various sources is generated to support the meaning assigned to assessment scores. The source of particular interest is “internal structure”, which relates to psychometric characteristics (i.e., difficulty and discriminatory ability) of multiple-choice questions.¹⁹ For instance, scores on an exam or sets of items intended to measure the knowledge of similar content should be highly correlated with each other. Such high correlation is best observed when items in an exam are of moderate difficulty (difficulty index = 0.4 - 0.8) and sufficient discriminatory ability (point biserial correlation ≥ 0.2).²⁰ When the majority (but not all) of the items in an exam are of moderate difficulty and sufficient discriminatory ability, a higher level of inter-item correlation is observed that

serves as evidence of sound internal structure of the exam.¹⁷⁻²⁰ It is worth noting that not all items in an exam need to be of moderate difficulty and sufficient discriminatory ability. For example, if a topic is important and taught and learned well during the administration of a curriculum, performance on an item assessing that topic's knowledge may be near perfect. Such an item may not exhibit a moderate difficulty index value, sufficient discriminatory ability, or high distractor functioning. However, it is important to include a few such items in an exam in order to assess knowledge of educationally important topics (such as benign vs. malignant nature of a breast lump), even if performance on such topics adds little value to the internal structure of the exam and the rank ordering of examinees.

Many of the statistical analyses needed to support or refute evidence of an exam's internal structure are often carried out as routine quality-control procedures in in-house exams. One of the most commonly used analyses is "item analysis", which computes each item's difficulty index, discriminatory index (any relevant index that shows how well that item separates high performing from low performing examinees), and selection frequencies of each option of an item. Overall summary statistics for an exam are also computed in commercially available item analysis software packages; the summary statistics show mean item difficulty index on an exam, mean discrimination index, as well as the reliability coefficient of scores obtained from the exam. The difficulty and discrimination indices used in this study are explained further in the "Data analysis" subsection of the Methods section of this chapter.

Materials and Methods

Research Design

A repeated-measures experimental research design was used. The study protocol was approved, and exempted from full review by the Institutional Review Board of University of North Dakota.

Subjects

Two cohorts of Year 1 medical students (Cohort 1 n = 69, Cohort 2 n = 70) at the University of North Dakota School of Medicine and Health Sciences from the graduating class of 2016 and 2017 served as subjects. Table III-2 displays gender representation (percentage of male and female students), grade point average (GPA) in undergraduate studies, and average medical college admissions test (MCAT) scores in each cohort. Some degree of variation in all these characteristics was seen across the cohorts, which may represent fluctuation in trends and institutional policies in regards to medical school admissions.

Table III-2. Demographic characteristics in each cohort of subjects.

	Gender representation (% of males, % of females)	Average undergraduate (pre-matriculation) GPA	Average Medical College Admissions Test score
Class of 2016	58.4%, 41.6%	3.71	28.0
Class of 2017	52.9%, 47.1%	3.71	27.8

The school's Patient-Centered Learning curriculum emphasizes social determinants of health and disease, as well as early application of scientific knowledge to patient care through lecture, laboratory, small group problem-based as well as simulation learning experiences. Multiple-choice exams are used among a number of methods to assess student learning at the end of each of the four 8-week curricular blocks in Year 1.

Assessment methods are criterion-referenced; a student must score 75% or better on an

end-of-block multiple-choice examination to have their performance interpreted as “satisfactory”.

Procedure

Fifty-five (55) multiple-choice questions with either too high difficulty (difficulty index <0.4), too low difficulty (difficulty index >0.8) or insufficient discriminatory ability (point biserial correlation coefficient <0.2) were identified from each end-of-block (Blocks I-IV) multiple-choice exam administered in the previous academic year. The psychometrics literature, discussed in detail in the “Introduction” section of this chapter, was used as a guide to select these criteria for multiple-choice difficulty and discriminatory ability.^{17–20} These items represented a variety of preclinical subjects including gross anatomy, physiology, biochemistry, pharmacology, genetics, developmental biology, and neuroscience.

Intervention

The design of the study involved random placement of questions (rather than subjects) in the experimental or the control group.

Twenty-one (21) questions were randomly placed in *Experimental Subgroup A*. These items underwent correction of item writing flaws (IWFs) along with enhancement of cognitive level (where needed) tested by the item via addition of a clinical or laboratory vignette.

First part of intervention in Experimental Subgroup A items was enhancement of cognitive level (if needed) tested by the item. Miller’s pyramid²¹ was used as a guide to enhance the cognitive level (CL) tested by the item. Simply put, inclusion of the clinical or laboratory vignette allowed assessment of the topic of interest through application of

knowledge and not just via plain factual recall. Inclusion of a clinical or laboratory vignette to enhance the cognitive level tested by the item was employed only on items that were previously testing plain factual recall.

The second part of the intervention in Experimental Subgroup A items was removal of any item writing flaws (IWFs) from the item. Guidelines published by National Board of Medical Examiners³ were used to identify and correct the item writing flaws. Table III-1 (in the “Introduction” section) displays the list of published item writing flaws with corresponding numerical codes utilized in this study. Table III-3 displays an example of an item that underwent this intervention, with step-by-step elaboration on the intervention itself.

Table III-3. Example of interventions in Experiment Subgroup A (Removal of IWFs and enhancement of tested cognitive level). Text in bold-italics elaborates on the step-by-step process of the intervention.

Before	After
<p>Which of the following best describes the location of the prostate gland? A: Inferior and posterior to the neck of the bladder in the rectovesical pouch B: At the neck of the bladder superior to the pelvic diaphragm** C: At the neck of the bladder inferior to the pelvic diaphragm D: In the superficial perineal pouch E: In the deep perineal pouch</p> <p><i>The topic of interest in this item was “location of the prostate gland”.</i></p> <p><i>The item was found to be testing low cognitive level owing to plain recall of a fact (i.e., location of the prostate gland).</i></p> <p><i>Moreover, the flaws identified in this item were:</i> <i>a. Long or complicated options</i> <i>b. Non-logical order of options</i></p>	<p><i>The author of this dissertation studied the topic “location of the prostate gland” from recommended textual references. A clinical vignette was developed to assess the same topic at a higher cognitive level and included in the revised version of the item upon the approval of the item’s original author.</i> A 72 years old male, in relatively good health, complains of frequent urination, weak stream, and post-void feeling of residual urine. Digital rectal exam reveals an enlarged organ. Which of the following describes the location of this organ?</p> <p>A: Deep perineal pouch B: Inferior to pelvic diaphragm C: Rectovesical pouch D: Superficial perineal pouch E: Superior to pelvic diaphragm**</p> <p><i>Item flaw “long or complicated options” was removed by simplifying the incorrect options (distractors). Note that the distractors underwent only simplification, and not removal or replacement with a different distractor.</i></p> <p><i>Item flaw “non-logical order of options” was removed by arranging the options in an alphabetical order.</i></p> <p><i>The revised version of the item was administered in end-of-curricular-block exam.</i></p>

Eleven (11) items were randomly placed in *Experimental Subgroup B*. Nine of these eleven items underwent replacement, and two of these items underwent removal, of the at least one non-functioning distractor. The removed or replaced distractors had demonstrated <5% selection frequency in the administration of the item in its previous administration in the end-of-block exam. Input from the faculty members who originally wrote these items was solicited throughout the intervention in this subgroup as well. For example, a distractor with <5% selection frequency was first identified through item analysis data. Then, the author of dissertation studied the topic under assessment from recommended texts. A list of other plausible distractors was created based on that reading, and the list was shared with the item's original author. Final decision of replacement or removal of distractors was left up to original author of the item, upon whose approval the revised version of the item was re-administered in the end-of-curricular block exam. It was decided, in advance, to assign fewer items to this category because some faculty expressed concern or unwillingness to replace or remove distractors. Also, the institutional policy required usage of items that have no fewer than four and no more than five options in in-house high-stakes exams. Table III-4 displays example of an item that underwent this intervention, with step-by-step elaboration on the intervention itself.

Twenty-three (23) items were blindly placed in the *Control group*. These did not undergo any intervention, and were re-administered *as-is* in the four high stakes end-of-block exams over the next academic year.

Table III-4. Example of interventions in Experiment Subgroup B (Replacement of non-functioning distractors). Text in bold-italics elaborates on the step-by-step process of the intervention.

Before	After
<p>A very premature infant is administered oxygen in the neonatal intensive care unit. Knowing that premature infants can also be cysteine-deficient, the patient is also given supplements of this amino acid to combat oxidative damage associated with oxygen toxicity. Cysteine is therapeutic because it is a precursor for what important intracellular antioxidant?</p> <p>A: Carnitine B: Glutathione** C: Histamine D: Phosphocreatine E: Serotonin</p> <p><i>The topic of interest in this item is “Therapeutic basis of cysteine as an intracellular antioxidant”.</i></p> <p><i>Four distractors (A, C, D, E) in this item showed <5% selection frequencies (non-functioning distractors).</i></p>	<p>A very premature infant is administered oxygen in the neonatal intensive care unit. Knowing that premature infants can also be cysteine-deficient, the patient is also given supplements of associated with oxygen toxicity. Cysteine is therapeutic because it is a precursor for what important intracellular antioxidant?</p> <p><i>The author of this dissertation studied the topic “Therapeutic basis of cysteine as an intracellular antioxidant” from recommended textual references. Four new distractors shown below (bold-italicized) were developed based on that reading.</i></p> <p><i>A: Melatonin</i> <i>B: Glutathione**</i> <i>C: Uric acid</i> <i>D: Vitamin C</i> <i>E: Vitamin E</i></p> <p><i>These new distractors were discussed with the item’s original author. The author agreed to use the new distractors as a replacement for the previously non-functioning distractors and to re-administer the revised version of the item in the end-of-curricular block exam.</i></p>

Data Collection and Analysis

The following data was collected for each item in the study.

- a. Psychometric characteristics, i.e. item difficulty index (diff.) and discriminatory ability (point biserial correlation coefficient; pbi), before and after revision of the item. Difficulty index, is defined as the proportion of test-takers answering the item correctly and is calculated as follows: number of correct answers / number of all answers. According to the literature,²² an item is classified as “moderately difficult” when its difficulty index lies between 0.4 and 0.8. Therefore, for the purpose of this study, items with difficulty index of >0.8 were classified as “too easy”, and those with difficulty index of <0.4 were classified as “too difficult”. Point biserial correlation (a.k.a. item-total correlation) is the correlation coefficient of scores on an item

with the total of scores on all other items in an exam. Point biserial correlation is an index of discriminatory ability of an item, and ranges from – 1.00 to + 1.00 with higher values indicating that performance on an item correlates well with the total score. If an item has low point biserial correlation (usually < 0.2), it is considered less helpful in separating high and low performing students and can be flagged for revision or removal from the exam.^{20, 22}

- b. The number of functioning distractors (FDs) in the overall exam as well as per multiple-choice question, before and after revision of each item. As discussed above, the definition used for a “functioning” distractor was “an incorrect option displaying $\geq 5\%$ selection frequency.”^{14, 15}
- c. Item-writing flaws (IWFs) in each item (Table III-1).
- d. Cognitive level (CL) (1 = low level, i.e., plain factual recall; 2 = high level, i.e., application of knowledge) tested by each item. Only two scales (1 and 2) were used to categorize the cognitive level tested by an item.

The collected data were stored in Microsoft Excel (2010) and analyzed via Microsoft Excel and SigmaStat v. 20.

Results

Experimental Subgroup A

Table III-5 shows flaw type, tested cognitive level (CL), number of functioning distractors (FDs), difficulty index (diff.) and point biserial correlation (pbi) before and after revision of each item in this subgroup. Table III-8 presents a summary of these results. Along with removal of identified item-writing flaws, 14 of these 21 items also

Table III-5. Flaw type, tested cognitive level (CL), # of functioning distractors (FDs), difficulty index (diff.) and point biserial correlation (pbi) before and after intervention in Experimental Subgroup A (IWF removal + enhancement of tested CL). CL: 1 = low / plain recall, 2 = high / application of knowledge.

Item ID	Flaw type	CL	Total distractors	FDs before	FDs after	Diff. before	Diff. after	pbi before	pbi after
1	5; 11	2	4	0	0	0.95	0.93	-0.03	0.19
2	11	2	4	0	0	1	1	0	0
3	11	2	4	0	0	0.96	0.97	-0.08	0.06
4	11	1	4	2	0	0.86	0.99	0.08	-0.04
5	11	2	4	1	1	0.91	0.85	0.16	0.49
6	None	1	4	1	0	0.87	0.96	0.08	0.17
7	11	1	4	0	1	0.93	0.77	0.02	0.08
8	11	2	4	1	1	0.89	0.93	0.12	0.29
9	11	1	4	1	1	0.91	0.84	0.06	0.38
10	5; 11	1	3	0	0	0.96	0.94	0.09	0.31
11	11	1	3	0	0	1	1	0	0
12	11	1	3	0	1	0.95	0.93	0.05	0.16
13	7; 13	1	4	2	3	0.55	0.43	0.08	0.17
14	None	1	3	0	0	0.93	0.97	0.09	0.2
15	13	1	3	1	1	0.66	0.45	-0.06	0.01
16	7; 11	1	4	2	3	0.66	0.68	-0.02	0.22
17	7; 11	1	3	0	0	1	1	0	0
18	11	1	3	2	3	0.55	0.5	0.12	0.23
19	11	1	4	0	1	1	0.87	0	0.39
20	11; 13	2	4	1	0	0.6	0.96	0.16	0.31
21	5; 7; 11; 13	2	3	0	1	0.97	0.87	0.13	0.39

underwent revision to enhance the cognitive level tested by the item (Table III-5, Column 3, CL = 1). After revision of the items in this subgroup, the average number of functioning distractors increased from 0.67 to 0.81, average item difficulty increased from 0.86 to 0.85 (1% increase), and average point biserial correlation increased from 0.05 to 0.19 (Table III-8). Collectively, in these 21 items, the number of items with moderate difficulty remained unchanged from 5, while the number of items with sufficient discriminatory ability increased from 0 to 10 (47% increase). The increase in the number of items with sufficient point biserial correlation from “0” before revision to “10” after revision was found to be statistically significant (Fisher’s Exact, $p = 0.000$). In short, removal of item flaws and enhancement of cognitive level (where needed) tested

by the items raised the collective discriminatory ability of this subgroup of items without much impact on their collective difficulty.

Experimental Subgroup B

Table III-6 shows flaw type, tested cognitive level (CL), number of functioning distractors (FDs), difficulty index (diff.) and point biserial correlation (pbi) before and after revision of each item in this subgroup. Table III-8 presents a summary of these results. The two questions that underwent removal, rather than replacement, of at least one non-functioning distractor are highlighted (shaded) in Column 1 of Table III-6. After revision of the items in this subgroup, the average number of functioning distractors increased from 0.90 to 1.09, average item difficulty index increased from 0.85 to 0.80 (5% increase), and average point biserial correlation increased from 0.04 to 0.19 (Table III-8). Collectively, in these 11 items, the number of items with moderate difficulty increased from 3 to 4 (9% increase), while the number of items with sufficient discriminatory ability increased from 0 to 6 (56% increase). The increase in the number of items with sufficient point biserial correlation from “0” before revision to “6” after

Table III-6. Flaw type, tested cognitive level (CL), # of functioning distractors (FDs), difficulty index (diff.) and point biserial correlation (pbi) before and after intervention in Experimental Subgroup B (IWF removal + enhancement of tested CL). CL: 1 = low / plain recall, 2 = high / application of knowledge.

Item ID	Flaw type	C L	Total distractors before	Total distractors after	FDs before	FDs after	Diff. before	Diff. After	pbi before	pbi after
22	5, 11	2	3	3	2	3	0.61	0.56	0.11	0.14
23	7; 11; 13	2	4	4	2	2	0.61	0.44	0.18	0.3
24	11	1	4	4	1	0	0.86	0.93	-0.03	0.16
25	4; 7; 11	2	3	3	0	1	1	0.81	0	0.3
26	11	2	4	4	0	0	0.95	0.97	0.11	-0.01
27	11	2	4	4	2	1	0.87	0.9	0.05	0.25
28	2, 11	1	4	4	0	1	0.97	0.91	-0.1	0.01
29	11	1	3	3	0	0	1	1	0	0
30	None	2	4	4	0	1	0.96	0.7	0.06	0.28
31	11	2	4	3	2	2	0.68	0.79	0.04	0.41
32	11	1	4	3	2	1	0.87	0.81	0.02	0.25

revision was found to be statistically significant (Fisher's Exact, $p = 0.012$). In short, replacement or removal of at least one non-functioning distractor was found to raise the collective discriminatory ability of this group of multiple-choice questions with a small impact on their collective difficulty level.

Control group (C)

Table III-7 shows flaw type, tested cognitive level (CL), number of functioning distractors (FDs), difficulty index (diff.) and point biserial correlation (pbi) upon initial administration as well as re-administration of each item in this subgroup. The summary of these results is presented in Table III-8. Upon re-administration, the average number of functioning distractors increased from 1.00 to 0.96, average item difficulty index increased from 0.84 to 0.83 (1% increase), and average point biserial correlation increased from 0.05 to 0.06 (Table III-8). Collectively, in these 23 items, the number of items with moderate difficulty increased from 10 to 12 (8% increase), while the number of items with sufficient discriminatory ability remained unchanged from 0. The increase in the number of items with moderate difficulty from "10" before revision to "12" after revision was found to be statistically insignificant (Fisher's Exact, $p = 0.763$). In short, small decrease in the average number of functioning distractors, and small increase in average difficulty and discriminatory ability was observed in this group of items upon re-administration without any intervention. These small changes perhaps reflect slight year-to-year fluctuation in performance on multiple-choice exams.

It is worth noting that similar flaws were discovered in the experimental and control group of items. Refer to Table III-1 for published item-writing flaws and their

Table III-7. Flaw type, tested cognitive level (CL), # of functioning distractors (FDs) and psychometric characteristics before and after intervention in Control group (C no intervention). CL: 1 = low / plain recall, 2 = high / application of knowledge.

Item ID	Flaw type	C L	Total distractors	FDs before	FDs after	Diff. before	Diff. After	pbi before	pbi after
33	2; 11	1	4	1	1	0.68	0.65	-0.06	0.01
34	5; 11	1	3	2	3	0.61	0.56	0.11	0.14
35	11	1	4	1	0	0.86	0.93	-0.03	0.16
36	11	1	4	0	0	0.98	0.97	-0.02	0.09
37	4	1	3	3	2	0.58	0.68	0	0.18
38	11; 13	2	3	1	1	0.56	0.54	0.01	0.09
39	11; 13	1	3	2	2	0.58	0.58	0.12	-0.07
40	11	1	3	0	1	0.98	0.89	0.01	0.07
41		1	4	1	1	0.89	0.9	0.13	0.05
42	11	1	3	1	1	0.85	0.9	0.11	0.14
43	11	1	3	1	0	0.88	0.9	0.15	0.13
44	11	1	3	0	0	0.97	0.97	0.1	-0.02
45	11	1	4	1	1	0.92	0.93	0.01	-0.07
46	11; 12	1	3	0	0	0.97	0.99	0.1	-0.06
47	11	1	4	2	2	0.72	0.68	0.15	0.18
48	11	1	4	1	1	0.95	0.91	0.15	0.17
49		1	4	0	0	1	0.97	0	-0.07
50	11	1	4	2	2	0.74	0.75	0.05	0.14
51	5	2	3	1	1	0.82	0.78	0.04	0.01
52	4; 11	1	4	0	0	1	1	0	0
53	11	1	4	2	2	0.82	0.79	-0.07	0.03
54	11	1	3	0	0	1	1	0	0
55	11	1	3	1	1	0.92	0.86	0.15	0.14

Table III-8. Summary of psychometric characteristics before and after intervention in experiment and control group, as well as the result of Fisher's exact analysis (highlighted cells). IWF: Item-Writing Flaws. CL: Cognitive Level. NFDs: Non-functioning distractors.

	Subgroup A (removal of IWFs + enhancement of CL)		Subgroup B (replacement or removal of NFDs)		Control Group	
	Before	After	Before	After	Before	Before
# of items	21	21	11	11	23	23
Ave. # of distractors per MCQ	3.62	3.62	3.73	3.55	3.48	3.48
Total # of distractors	76	76	41	39	80	80
Total # of FDs	14 (18%)	17(22%)	10 (27%)	12 (33%)	23 (29%)	22 (28%)
Mean # of FDs	0.67	0.81	0.91	1.09	1	0.96
Ave. Diff.	0.86	0.85	0.85	0.8	0.84	0.83
Ave. pbi	0.05	0.19	0.04	0.19	0.05	0.06
# of MCQs with moderate difficulty	5	5	3	4 (9% increase; df [1], p > 0.05)	10	12 (8% increase; df [1], p > 0.05)
# of MCQs with sufficient discriminatory ability	0	10 (47% increase; df [1], p < 0.05)	0	6 (56% increase; df [1], p < 0.05)	0	0

corresponding codes used in this study. A glance at Tables III-5, III-6 and III-7 reveals that the most common flaws in each group were non-logical order of options (Flaw# 11), long, complicated or double options (Flaw # 7) and word-repeats between the stem and correct answer (Flaw# 5). Other less common flaws seen in each group were tricky or unnecessarily complicated stems (Flaw# 13), collectively exhaustive subset of options (Flaw# 2), and long correct answer (Flaw# 4).

However, the level of tested cognitive function in the experimental and control group of items showed dissimilarity. In Experimental Subgroup A, 14 out of 21 total items (67%) were found to be testing low cognitive level (Table III-5, column 3). All of these 14 items underwent revision to enhance the tested cognitive level via incorporation of a clinical or laboratory vignette. On the other hand, in Experimental Subgroup B, 36% items (4 out of 11) were found to be testing low cognitive level (Table III-6, column 3), and in Control group C, 91% items (21 out of 23) were found to be testing low cognitive level (Table III-7, column 3). None of these items in the Experimental Subgroup B and Control group C underwent revision to enhance the tested cognitive level since this revision did not apply to these groups of items. Since fewer Experimental Subgroup B items (36%) were originally written at low cognitive level than the Experimental Subgroup A (61%) and Control Group C (91%) items, these groups were not equivalent in this regard. The possible effect of this lack of equivalency on the outcome of interest (item difficulty and discriminatory abilities post-revision) is brought up in the “Discussion” section of this chapter.

Discussion

The experimental study presented in this chapter addressed the impact of item flaws, testing of low cognitive function and low distractor functioning on the quality of multiple-choice questions used in in-house exams in Year 1 preclinical medical education. Since item quality is closely related to validity and reliability of assessment,^{19, 23} the faculty and personnel responsible for designing high-stakes assessment in undergraduate medical education strive for optimal quality of items used in in-house exams. However, these faculty and personnel may face difficulty in accomplishing this goal owing to lack of formal training in item writing and lack of awareness of item quality parameters, since their typical role is to teach basic science content³. Therefore, the research question raised and findings presented in this study may have relevance to a broad group of medical educator scholars interested in raising and maintaining the quality of their in-house assessment. A few thoughts on the obtained results are shared below.

Firstly, item flaw correction (along with enhancement of tested cognitive level) (Experimental Subgroup A) and replacement or removal of non-functioning distractors (Experimental Subgroup B) increases the number of functioning distractors ($\geq 5\%$ selection frequency) per item to a similar degree. Flaw removal (along with enhancement of tested cognitive level) led to an *increase* in number of functioning distractors per item from 0.67 to 0.81, while replacement or removal of non-functioning distractors led to an *increase* in number of functioning distractors per item from 0.91 to 1.09 (Table III-8). On the other hand, a slight *decrease* in the number of functioning distractors per item from 1 to 0.96 was noted in the control group of items. This shows that flaw removal

(along with enhancement of tested cognitive level) and replacement or removal of non-functioning distractors helps in enhancing distractor functioning to some degree. This outcome provides some leverage in gauging examinees' conceptual misunderstandings through multiple-choice questions.

However, the level of increase in the number of functioning distractors per item was still not as high as the investigators in this study expected from either intervention. The high stakes end-of-curricular block exams in Year 1 medical education at University of North Dakota School of Medicine and Health Sciences comprise items with either four or five total options. Our desire was to see a greater increase in the number of distractors with $\geq 5\%$ selection frequency (*functioning* distractors) as an outcome of item flaw removal (along with enhancement of tested cognitive level) (Experimental Subgroup A) as well as of replacement or removal of non-functioning distractors (Experimental Subgroup B). However, the numbers of 0.81 (from 0.67; Experimental Subgroup A) and 1.09 (from 0.91; Experimental Subgroup B) average functioning distractors per item were seen as the outcome of revisions in these groups (Table III-8). These numbers are not high considering the 3 or 4 distractors (in 4- or 5- option item) provided per item. Perhaps, developing more plausible distractors from examinee responses on free-response (fill-in-the-blank) version of the items could help in this regard. The study presented in Chapter 2 of this dissertation used this method for items assessing knowledge of neurohistology with promising results. An expansion of that design will help make definite conclusions in this regard.

It is worth noting that the improvement in number of functioning distractors per item noted from our interventions (0.81 from 0.67, Experimental Subgroup A; 1.09 from

0.91, Experimental Subgroup B) was similar to a study published by Tarrant et al. in 2010¹⁶. In their study, 41 4-option MCQs were converted to 3-option ones via removal of the least functioning distractor; upon re-administration, the 3-option version of the items experienced an increase in the number of functioning distractors per item from 1.32 to 1.49.¹⁶ This shows that a small increase in number of functioning distractors per item is not an unlikely outcome of the revisions used in our study. Perhaps, the lack of any significant finding in this regard hints at the limited value of our revision. Maybe, flaw removal along with enhancement of tested cognitive level, and replacement or removal of non-functioning distractors have limited value when performed individually on any item. Greater summative enhancement in distractor functioning may be noted if *both* interventions were performed simultaneously on each item; this is a research question for a future study.

Secondly, item flaw removal (along with enhancement of tested cognitive level) has less of an impact on average item difficulty (average 1% increase) when compared to replacement or removal of non-functioning distractors (average 5% increase) (Table III-8). Therefore, in order to construct optimally difficult criterion-referenced examinations,²⁰ item writers may focus more on replacement of non-functioning distractors with more plausible ones. Tarrant and Ware, in their 2010 study, concluded similarly on the role of distractor functioning in construction of optimally difficult exams.¹⁶ Their study involved removal of the least functioning distractor from 4-option multiple-choice questions. They reported a 3% increase (0.70 from 0.73) in average item difficulty as an outcome of a small rise in the number of functioning distractors per item

(1.49 from 1.32)¹⁶. This shows that replacement or removal of non-functioning distractors raises item difficulty, albeit to a slight extent.

Thirdly, both interventions improve the ability of multiple-choice questions to discriminate among high and low ability students. An average point biserial correlation of 0.19 was observed with both the removal of item flaws (along with enhancement of tested cognitive level) (Experimental Subgroup A), and replacement or removal of non-functioning distractors (Experimental Subgroup B) (Table III-8). This value (0.19) approximates the sufficient point biserial correlation coefficient recommended for in-house assessment (0.20) according to the published medical education literature.²⁰ Moreover, both interventions resulted in an increase in the number of items with sufficient point biserial correlation coefficients (47% and 54% in Experimental Subgroup A and B, respectively), while no change was observed upon item re-administration of the control group of items (Table III-8). This affirms comparable utility of each revision for better discrimination among high and low performing students. This finding is in sync with the Tarrant and Ware study published in 2010.¹⁶ They reported that after removal of the least functioning distractor from 41 4-option MCQs, fewer 3-option items exhibited poor discriminatory ability, and discriminatory ability of the remaining distractors was found to be increased.¹⁶ Better average discriminatory ability seen post-intervention in our study allows for identification of students with true conceptual misunderstandings, which can be addressed through remediation and modification of learning strategies. Moreover, better average discriminatory ability implies that performance on any item is less influenced by factors other than the student's knowledge of the content area. Such factors include ambiguous or confusing stems or options, whose presence introduces

more difficulty to the item without any connection with the topic under assessment.²³ The variance in performance resulting from such irrelevant difficulty weakens the validity evidence of scores obtained on high-stakes exams. Downing and Haladyna have discussed this phenomenon in a seminal 2004 paper published in the journal *Medical Education*.²³ They describe the performance variance introduced by factors such as ambiguous or confusion options or stems as “Construct-Irrelevant Variance”, and elaborate on how such factors systematically interfere with meaningful interpretation of scores obtained on high stakes exams. Overall, the paper by Downing and Haladyna aptly discussed the impact of item flaws on the quality of multiple-choice assessment.

A few limitations to the study’s findings exist. Firstly, the number of items used in this study was small, especially in the Experimental Subgroup B. Therefore, obtained results should be generalized with caution, especially of the impact of replacement or removal of non-functioning distractors. Expansion of this study to perform revisions, where necessary, on all items in an exam, as well as replication of this experimental design by scholars located elsewhere will further strengthen the case for utility of these revisions. Secondly, the level of tested cognitive function in the experimental and control groups were dissimilar. Among the items that underwent removal of flaws along with enhancement of tested cognitive level (Experimental Subgroup A), 14 out of 23 total items (61%) were found to be testing low cognitive level pre-intervention (Table III-5, column 3). While among the items that underwent removal of non-functioning distractors (Experimental Subgroup B), 4 out of 11 total items (36%) were found to testing low cognitive level (Table III-6, column 3). In the Control group, 21 out of 23 total items (91%) were discovered to testing low cognitive level (Table III-7, column 3).

The fact that, before revision, the number of items in the experimental and control groups were not equal in terms of tested cognitive level may account for some of the differences in difficulty and discriminatory ability after revision. This disparity may have confounded our results. In the expansion of the study, we aim to ensure items in each group are at similar cognitive levels so that any improvement in outcomes could be attributed more confidently to our intervention. Thirdly, neither removal of item flaws (along with enhancement of tested cognitive level), nor replacement or removal of non-functioning distractors resulted in a significant change in item difficulty in our study. One of the goals of these interventions was to bring the average item difficulty within a range of 0.4 – 0.8; such a moderate range of item difficulty helps discriminate more adequately amongst high- and low-performing examinees and improves the reliability of obtained scores as well.²⁰ Despite falling short of this goal, the revisions still brought improvement in discriminatory ability of our experimental group of items; this finding is promising since it enhances the internal structure validity evidence of obtained examination scores.^{19,20}

In conclusion, the use of item analysis data in evaluating the quality of multiple-choice exams and an understanding of the role of quality item construction and distractor functioning is necessary to effectively assess student learning on high-stakes in-house exams. From the authors' experience, focusing on item flaws, cognitive level tested by the items and distractor functioning requires vigilance, skill and resources. Moreover, faculty professional development in this area can be a challenging task. However, the data demonstrate that the outcome is worth the effort, i.e., in-house exams that reliably

identify truly competent learners who should progress to the next stage of training, as well as those who may require remediation.

References

1. Mehrens WA, Lehmann IJ. Measurement and Evaluation in Education and Psychology. Fort Worth, TX. Holt, Rinehart and Winston, Fort Worth TX. 1991.
2. Masters JC, Hulsmeyer BS, Pike ME, Leichty K, Miller M.T, Verst AL. Assessment of multiple-choice questions in selected test banks accompanying textbooks used in nursing education. J Nurs Educ. 2001;40(1);25–32.
3. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew H. The quality of in-house medical school examinations. Acad Med. 2002;77:156–161.
4. Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. National Board of Medical Examiners. Philadelphia, PA. [Internet].2011 Nov 18 ; Available from <http://www.nbme.org/publications/item-writing-manual-download.html>
5. Downing SM. 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. Adv Health Sci Educ Theory Pract. 2005;10:133–143.
6. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ. 2008;42:198–206.
7. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002;15(3):309–334.
8. Tarrant M, Knierim A, Hayes S K, Ware J. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurs Educ Today. 26(8):662–671.

9. Newble D. A comparison of multiple-choice and free-response tests in examination of clinical competence. *Med Educ.* 1979;13:263–268.
10. Maguire T, Shakun E, Harley C. Setting standards for multiple-choice items in clinical reasoning. *Eval Health Prof.* 1992;15:434–452.
11. Elstein A. Beyond multiple-choice questions and essays: the need for a new way to assess clinical competence. *Acad Med.* 1993;68:244–249.
12. Boshuizen H, van der Vleuten C, Schmidt H, Machiels-Bongaerts M. Measuring knowledge and clinical reasoning skills in a problem-based curriculum. *Med Educ.* 1997;31:115–121.
13. Shakun E, Maguire T, Cook D. Strategy choices in multiple-choice items. *Acad Med.* 1994;(10 suppl):S7–S9.
14. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ.* 2004;38:327–33.
15. Rodriguez, MC. Three options are optimal for multiple-choice items: a meta analysis of 80 years of research. *Educ Measure Issues Prac.* 2005;24(2):3–13.
16. Tarrant M, Ware J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. *Nurs Educ Today.* 2010;30(6):539–543.
17. Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Measure Issues Prac.* 1995;14:5–8.
18. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ.* 2003;37:830–837.

19. Kern DE, Thomas PA, Hughes MT. Curriculum development for medical education: A six-step approach, 2nd edit. Baltimore, MD: Johns Hopkins University Press; 2009.
20. De Champlain A. A primer on classical test theory and item response theory for assessment in medical education. Med Educ. 2010;44:109–117.
21. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65: 563–567.
22. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33:447–458.
23. Downing SM, Haladyna TM. Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38:327–333.

CHAPTER IV

SHORT- AND LONG-TERM RETENTION OF KNOWLEDGE OF HUMAN ANATOMY – THE ROLE OF RETRIEVAL PRACTICE

Abstract

Introduction

Repeated retrieval practice, in the form of tests or quizzes, has been known to enhance learners' ability to retain knowledge. Free-response (FR) questions have been reported to be more useful in this regard than multiple-choice questions (MCQs). Most previous reports of enhanced knowledge retention were based on usage of the *same* items in repeated testing as well as final assessment on a topic. The study presented here determines whether knowledge retention is enhanced via usage of *different* items in repeated testing on a topic. Specifically, the impact of format (FR vs. MCQ) and frequency (thrice vs. once) of retrieval practice on short-term (4 weeks) and long-term (2 – 7 months) retention of knowledge was studied.

Materials and Methods

A within-subjects experimental design was used. Sixteen (16) radiologic and twelve (12) non-radiologic anatomy topics across four curricular blocks in Year 1 medical education were identified and randomly placed in the Experimental or the Control group of topics. The experimental group comprised two subgroups. Experimental Subgroup A comprised topics that were tested thrice via FR (free-response) questions, while Experimental Subgroup B comprised topics that were tested thrice via

MCQs (multiple-choice questions). The control group (C) comprised topics that were tested once via MCQs. Testing occurred in gross anatomy laboratory during cadaver dissection hours in the form of no-stakes short quizzes and was conducted only in the first half of each 8-week curricular block. Short-term retention (4-weeks after the last in-lab testing session) was evaluated by comparing average performance on the in-lab tests with that on the end-of-curricular block exam given in MCQ format. Long-term retention (2 – 7 months after end-of-curricular block exam) of select topics from the first three curricular blocks was assessed via comparison of end-of-curricular block performance with performance on an end-of-year quiz given in MCQ format.

Results

In regards to *short-term* retention of *radiologic* anatomy content, three out of four Experimental Subgroup A topics exhibited 41% – 53% gain in retention, while performance on one topic exhibited 42% decline. Two out of three Experimental Subgroup B topics exhibited 8% – 35% gain in retention, while performance on one topic exhibited 12% decline. Eight out of nine control group (C) topics exhibited 5% – 47% gain in retention, while performance on one topic exhibited 11% decline. In regards to *short-term* retention of *non-radiologic* anatomy content, all three Experimental Subgroup A topics exhibited 51% – 73% gain in retention. Two out of five Experimental Subgroup B topics exhibited 28% – 55% gain in retention, while performance on the other three topics exhibited declines ranging between 2% - 41%. Performance on the four control group (C) topics exhibited 4% – 58% gain in retention.

In regards to *long-term* retention of *radiologic* anatomy content, all three Experimental Subgroup A topics exhibited declines ranging between 11% and 38%.

Similarly, performance on both Experimental Subgroup B topics exhibited declines ranging between 3% and 38%. In contrast with the Experimental group, performance on the three control group (C) topics exhibited 2% – 4% gain in long-term retention. In regards to *long-term* retention of *non-radiologic* anatomy content, the two Experimental Subgroup A topics exhibited 10% – 49% decline in retention. One out of three Experimental Subgroup B topics exhibited 48% gain in retention, while the other two exhibited 5% to 10% decline in long-term retention.

Conclusions

In line with published reports, short-term (4 weeks) retention of human anatomy content tested repeatedly (thrice) via free-response questions tends to be better than that of short-term retention of content tested repeatedly via multiple-choice questions. This finding may not translate to long-term (2 – 7 months) retention of radiologic anatomy content; instead, testing only once via multiple-choice questions may be of benefit in this regard. Overall, in-lab tests can be useful in self-assessment and feedback on the subject of anatomy in Year 1 medical education.

Introduction

Retention of knowledge can be enhanced through no-stakes repeated testing of the studied and taught material.¹ Such retention has been linked to both “indirect” and “direct” effects of testing.^{2,3} The “indirect” effect refers to improvement in study strategy and efficient time management resulting from frequent testing.^{2,3} On the other hand, the “direct” effect (a.k.a. the “testing” effect) is based on enhancement of neuronal connections pertaining to a specific memory and holds true across a variety of curricular content and experimental conditions.³ Investigations on the direct effect of testing on

knowledge retention and skill acquisition have been reported in literature. A brief synopsis of these studies follows.

A study published in the journal *Science* by Karpicke and Roediger reported learning of a foreign language vocabulary words among undergraduate psychology students at Washington University at St. Louis.¹ The subjects (students) began by studying a list of 40 Swahili-English word pairs (e.g., *mashua*-boat) in a study period, and then testing over the entire list in a test period (e.g., *mashua*-?). After a subject had correctly produced a vocabulary word-pair during the testing phase, the effect of three conditions on retention of that knowledge was analyzed. These conditions were, a. repeated testing but dropping the word-pair from further study, b. repeated studying but dropping the word-pair from further testing, and c. dropping the word-pair from both studying and testing. Knowledge retention was assessed via recall of word-pairs after 1 week. Recall of repeatedly tested word-pairs was much better than the recall of repeatedly studied word-pairs (Cohen's $d = 4.03$). Moreover, recall of the word-pairs dropped from further testing after successful recall was found to range from 10% to 60%, while that of word-pairs tested repeatedly even after successful recall was found to range from 63% to 95%. The study concluded that repeated testing may be beneficial for retention of learned content and discussed how repeated retrieval practice can consolidate learning in a variety of educational contexts.

A study published in the journal *Medical Education* by Krommen et al. investigated whether testing effect also applies to learning of skills.⁴ Specifically, they investigated whether testing as final activity in a skills course increased the learning outcome compared with an equal amount of time spent on just practicing the skill. The

study was conducted in the context of a 4-hour in-hospital resuscitation course that is run the seventh semester of medical education at University of Copenhagen, Denmark. A total of 140 medical students, who served as subjects, were randomized into either the intervention or the control group. The intervention group underwent 3.5 hours of instruction and training on resuscitation followed by 30 minutes of testing. The control group underwent 4 hours of instruction and training. Two weeks after the course, learning outcome (usage of acquired skills) was assessed through a simulated clinical encounter. A pre-developed checklist of essential resuscitation skills was used for grading by the instructors. Learning outcome was found to be significantly higher in the intervention group (mean score 82.8%, 95% confidence interval 79.4–86.2) compared with the control group (mean score 73.3%, 95% confidence interval 70.5–76.1) ($p < 0.001$). Effect size (Cohen's d) of the difference of scores between the intervention and control groups was found to be 0.93. This study demonstrated the feasibility of implementing testing as final activity in simulation-based medical training and also showed the usefulness of testing in retention of important procedural skills. The effect of testing in the broader context of in-hospital training of future physicians was also discussed.

In the field of postgraduate medical education, a study published in the journal *Medical Education* by Larsen et al. investigated whether repeated testing enhances final recall of content at a more educationally relevant interval of 6 months.⁵ Postgraduate medical trainees (residents) in Pediatrics and Emergency Medicine participated in an interactive teaching session on two topics, “status epilepticus” and “myasthenia gravis”. Then, the trainees were randomized to two groups, which either took tests on status

epilepticus and studied a review sheet on myasthenia gravis (SE-TMG-S group), or took tests on myasthenia gravis and studied a review sheet on status epilepticus (MG-T/SE-S group). The review sheets consisted of information identical to that on the answer sheets for the tests. Testing on both topics comprised free-response (short-answer) questions. Testing and studying occurred immediately after the first interactive teaching session and then at two additional times at 2-week intervals; each testing session was followed by a discussion session. A final test was taken at an interval of about 6 months on both topics. Nineteen trainees in the SE-T/MG-S group and twenty-one trainees in the MG-T/SE-S group completed the study. In the final test, repeated testing was found to produce an average 13% higher score than repeated studying (39% versus 26%) ($p < 0.001$). Effect size (Cohen's d) of the difference between scores obtained on repeatedly tested and repeatedly studied topics was 0.91. The study showed that repeated testing, combined with feedback, helps in long-term (6 months) retention of knowledge, among postgraduate trainees, than repeated studying. Usefulness of testing as a tool for learning, and not just for assessment, was discussed in the broader context of medical education.

Since the study presented in this chapter of the dissertation encompasses the subject of human anatomy, a study published in the journal *Advances in Physiology Education* by John Dobson is worthy of mention here.⁶ Dobson investigated whether retrieval practice improves retention of knowledge in an undergraduate anatomy and physiology course.⁶ Dobson also investigated whether there is any difference in the degree of retention if either an expanding or a uniform pattern of retrieval practice is followed. Subjects (undergraduate students enrolled in the "Anatomy and Physiology" course) were randomly assigned to groups that underwent repeated testing either on an

expanding (n = 46; 1-, 2- and 3-week) or a uniform (n = 45; 3-week) schedule. Each group completed a total of 10 retrieval quizzes. Another group of students, which did not undergo repeated retrieval practice, served as the control group (n = 143). Final retention of content was assessed through a comprehensive exam during the last week of the semester. No significant difference was found between scores obtained by groups that followed either the expanding or the uniform schedule of retrieval practice. However, both these groups performed better (41% higher average score) on the comprehensive exam than did the control group ($F = 129.8, p = 0.00$). Effect size of the difference among mean scores in each group was measured through “Eta-squared”, which is the standard measure of effect size for ANOVA, and was found to be 0.36. This study showed how retrieval practice could be an effective strategy for enhancing the retention of content of undergraduate anatomy and physiology courses.

While studies such as above have reported on the general benefit of retrieval practice in retention of knowledge, others have discussed the superiority of repeated production tests (e.g., free-response questions) over repeated recognition tests (e.g., multiple-choice questions) in this regard.^{2,3} The reason for enhanced benefit of repeated production tests is that they require greater retrieval effort and depth of mental processing than recognition tests.^{2,3} This phenomenon has been explained via the difference between storage strength (relative permanence) and retrieval strength (momentary accessibility) of a memory trace.⁷ It is suggested that more effortful retrieval practice (such as via production tests) enhances storage strength to a much greater extent than easier retrieval practice (such as via recognition tests). Such an enhancement in storage strength resulting from repeated production tests leads to a deeper and longer-lasting

memory imprint.^{7,8} A study by Butler and Roediger published in the European Journal of Cognitive Psychology elaborated on this phenomenon.⁸ In their experiment, students enrolled in an undergraduate psychology course attended a series of three lectures on consecutive days. On each day, all students engaged in a different type of post-lecture activity such as taking a multiple-choice test or taking a free-response test. A final comprehensive test comprising free-response questions was given one month later to all students. Scores obtained on lecture material repeatedly tested via free-response questions were found to be higher than scores obtained on lecture material repeatedly tested via multiple-choice questions (mean proportion correct 0.47 and 0.36, respectively). A study by McDaniel et al. also found higher retention of knowledge via repeated free-response testing than repeated multiple-choice testing.⁹ Both these studies aptly demonstrate how production tests (such as free-response questions) can be superior to recognition tests (such as multiple-choice questions) in retention of learned content.

A closer look at the studies discussed above reveals that, in most studies, the *same* questions were used in repeated tests and final assessment of a topic. Therefore, the resulting enhanced knowledge retention can easily be understood in lieu of repeated processing of the same information. However, the study presented in this chapter of the dissertation investigates whether usage of *different* questions in repeated testing of a topic can enhance storage and retrieval strengths of that topic's memory. A literature search would show that a few studies have looked at this aspect of test-enhanced learning. For example, a study by Foos and Fisher assessed the value of test taking as a means of increasing learning among 105 undergraduate students.¹⁰ Students were given either an initial test or no test about the text material on the topic "American Civil War". The form

of the initial test was either fill-in (free-response) or multiple-choice, and the knowledge examined in the initial test was either directly stated in the reference reading (*verbatim*) or could be logically derived (*inferential*) from the reference reading. On the common final test, given two days later, retention of material tested initially by fill-in (free-response) questions was found to be greater than retention of material tested initially by multiple-choice questions. Moreover, retention of material tested via *inferential* questions was greater than retention of material tested via *verbatim* questions. It is worth noting that performance on inferential questions on a topic relies more heavily on understanding of the concept under assessment, while performance on *verbatim* questions on a topic may rely more heavily on rote memorization of a fact or the correct answer itself. The purpose behind Foos and Fisher's "inferential" questions was similar to the purpose behind our usage of *different* questions on repeated testing of a topic, i.e. to see whether repeated testing of a concept enhances its retention.

Interestingly, there is a dearth of studies evaluating the effect of repeated testing of a topic via *different* questions, especially in undergraduate medical education. The study by Foos and Fisher was conducted in the context of undergraduate history education. A literature search revealed just one such study conducted in the context of learning in anatomical sciences, which was published by Logan et al. in the journal *Anatomical Sciences Education*.¹¹ In their study, Logan et al. determined whether frequent quizzing had any effect on retention of knowledge in human anatomy. Short fill-in-the-blank (free-response) quizzes were given in a controlled setting to 21 undergraduate students aspiring to enter medical or dental schools. The quizzes were given on a weekly schedule and comprised questions on regional anatomy as well as the

nervous system. Each question on the nervous system was given three times, in a slightly different form each time. An example of a question's three forms, shared by Logan et al., follows.¹¹

Question on topic X given in Quiz 1: "Preganglionic cell bodies in the parasympathetic system are found at the _____ level of the spinal cord." (Answer: sacral)

Question on topic X given in Quiz 2: "The parasympathetic neurons at the sacral level of the spinal cord are _____-ganglionic." (Answer: pre-)

Question on topic X given in Quiz 3: "Preganglionic cell bodies at the sacral spinal level are a feature of the _____ division of the autonomic nervous system." (Answer: parasympathetic)

Logan et al. gave the second quiz approximately half an hour after the first one, and gave the third quiz one week after the second quiz. Average performance on the questions on repeatedly tested content was found to increase by almost 9% on the second quiz, and a further 20% on the third quiz (29% higher average score on the third quiz than on the first quiz). A final exam was given at the end of the semester. A positive correlation between performance on the quizzes and the final examination was found ($r(19) = 0.51, p < 0.02$). While the study by Logan et al. showed the value of repeated testing in learning of human anatomy content, its usage of *different* questions on a given topic was particularly interesting. The three forms of the example question shown above demonstrate how Logan et al. intentionally tested the same concept via three slightly different questions. The study presented in this chapter of the dissertation explores this avenue further.

The direct effect of testing via free-response and multiple-choice questions on retention of human anatomy educational content was investigated. The experimental group of topics underwent repeated (thrice) testing either via different multiple-choice questions on the same topic or different free-response questions on the same topic, while the control group of topics underwent testing only once via multiple-choice questions. The research question was: does the format (free-response vs. multiple-choice) and frequency (repeated vs. once) of tests using *different* questions on the same topic influence short- (4 weeks) and long-term (2 – 7 months) retention of anatomy knowledge? This avenue is of relevance in light of the complexities involved in clinical application of knowledge of human anatomy. Since such application requires expeditious recall of previously acquired knowledge, no-stakes testing may be a way to expedite and facilitate such recall. The overall aim of the study was to corroborate findings of other scholars in the context of Year 1 medical education and to evaluate the notion that repeated retrieval practice is beneficial for knowledge retention.

Materials and Methods

Subjects

One cohort of Year 1 medical students (the graduating class of 2016) at the University of North Dakota School of Medicine and Health Sciences served as subjects. This cohort (n = 69) had 58.4% male and 41.6% female students. Average undergraduate (pre-matriculation) grade point average (GPA) in undergraduate studies in this cohort was 3.171, and average medical college admissions test (MCAT) score in this cohort was 28.0.

The school's curriculum is a hybrid of Patient-Centered Learning (PCL) and discipline-based instruction. Participation in the study was voluntary; no consent was

sought, no personal or identifiable information was collected from the students and no points were granted for participation. Students were informed, in advance, of the nature of the study and self-assessment benefit of the in-lab tests. The study was approved by Institutional Review Board of University of North Dakota and was classified exempt from detailed review.

Materials

A within-subjects design was used. Sixteen (16) radiologic anatomy and twelve (12) non-radiologic anatomy (clinical anatomy) topics outlined in lecture and laboratory learning objectives were identified. The topics were blindly placed in two groups: experimental and control.

Experimental Subgroup A topics (four radiologic and three non-radiologic anatomy topics) were tested thrice via free-response items, while Experimental Subgroup B topics (three radiologic and five non-radiologic anatomy topics) were tested thrice via multiple-choice questions. Control group (C) topics (nine radiologic and four non-radiologic anatomy topics) were tested only once, via multiple-choice questions.

Most questions in all groups of topics required knowledge of one or two pieces of information. Figure IV-1 shows examples of questions used in repeated testing of Subgroup A (tested thrice free-response questions) and Subgroup B (tested thrice multiple-choice questions) topics.

Procedure

Frequency of testing was once per week for three consecutive weeks for the experiment group of topics, and only once for control group of topics. Testing was conducted only in the first half of each 8-week curricular block (total 4 blocks)



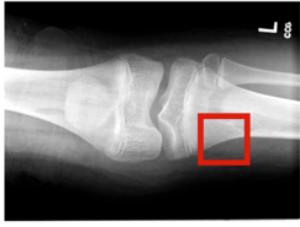
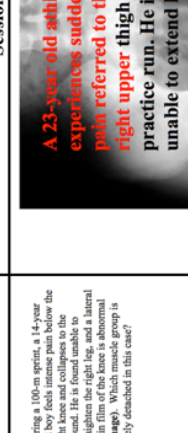
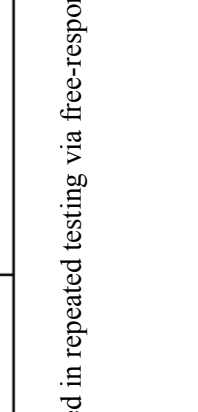
Session 1	Session 2	Session 3
<p>Subgroup A topic Anterior and medial thigh musculature</p>	<p>A 23-year old athlete experiences sudden, shooting pain referred to the back of his right upper thigh during a practice run. He is found unable to extend his thigh, and an AP radiograph is abnormal (image). Which group of muscles is likely to be detached in this case?</p> 	<p>An athlete presents with left knee pain. History reveals overuse of the thigh and leg musculature. There is pain and swelling at muscular insertion in the highlighted area (image). What is the nerve supply of the involved musculature?</p> 
<p>Subgroup B topic Gluteal and posterior thigh musculature</p>	<p>A 45 year old male with constant dull hip pain and weakness is found to have sarcoma of proximal femur. An AP plain film reveals a pathologic fracture (image). Which of the following muscles is most likely to be affected in this case?</p> 	<p>A 45 y.o. female presents with hip pain after slipping and falling on a wet floor. An oblique AP radiograph reveals fracture at a distinct site (image). Her right lower extremity appears slightly shorter and her right foot is laterally pointed. Which of the following muscles is most likely responsible for the abnormal positioning of her lower limb?</p> 
<p>A 68 year old woman fell in her bathroom and landed on her hip. An AP plain film of the hip revealed fracture in the arrowed area (image). Which of the following muscles is most likely to be detached in this case?</p>  <p>A. Biceps femoris B. Gluteus maximus C. Gluteus medius D. Iliopsoas E. Sartorius</p>	<p>A 45 year old male with constant dull hip pain and weakness is found to have sarcoma of proximal femur. An AP plain film reveals a pathologic fracture (image). Which of the following muscles is most likely to be affected in this case?</p> <p>A. Adductor magnus B. Biceps femoris C. Iliopsoas D. Rectus femoris E. Sartorius</p>	<p>A. Semitendinosus B. Adductor longus C. Gluteus maximus D. Rectus femoris E. Gracilis</p>

Figure IV- 1. Examples of questions used in repeated testing via free-response (Subgroup A topics) and multiple-choice (Subgroup B topics).

and inside the gross anatomy laboratory. Each in-lab test comprised 8 questions covering the experimental and control group of topics relevant to that curricular block. All in-lab testing sessions were conducted in the following manner.

- Halfway through each 2-hour laboratory session, cadaver dissection was momentarily halted and pre-developed answer sheets specific to that testing session were distributed.
- Questions, with any relevant images, were displayed one-by-one on large screen High-Definition TVs and one minute was provided to answer each question.
- At the end of each test, answer sheets were collected from all students, followed by an interactive discussion of each question.

A typical in-lab testing session lasted around 20 minutes. All answer sheets were coded and scored according to pre-developed answer keys. Performance data on each in-lab test was entered and stored in MS-Excel files.

Outcome Analysis

Short-term retention (four weeks) was assessed for all radiologic and non-radiologic anatomy topics. Short-term retention was assessed by comparing average performance (item difficulty) on the three in-lab tests with performance on the end of curricular block exam given in multiple-choice format four weeks later. Long-term retention (2 – 7 months) was assessed for nine blindly selected radiologic anatomy topics from the first three curricular blocks. Long-term retention was assessed by comparing end of curricular block performance with performance on an end of the year quiz. The intervals between end of the year quiz and end of curricular blocks 1, 2 and 3 exams were 7, 5 and 2 months respectively.

Effect size of any gains in retention in performance was calculated via Cohen's d . Effect size is an index of the degree to which finding of an experiment has practical significance in the study population regardless of the size of study sample.¹⁴ Cohen's d is a statistic that is equal to the difference between the means of experimental (M_e) and control (M_c) groups divided by the standard deviation of the control group (σ_c) (Cohen's $d = \frac{M_e - M_c}{\sigma_c}$).¹⁴ However, in cases where means being compared are from two measurements taken from the same group of subjects (our case), psychometric scholars recommend using the baseline standard deviation in the denominator of the formula.¹⁴ Hence, the following formula was used to calculate Cohen d in this study: $\frac{M_1 - M_2}{\sigma_1}$, where M_1 and M_2 represent mean performance (item difficulty) before and after intervention (in-lab testing). Statistical analyses were performed via MS-Excel and SigmaStat v. 20.

Results

Short-term Retention

Table IV-1 displays short-term retention (average in-lab item difficulty vs. end-of-block exam item difficulty) of radiologic anatomy topics. Three out of four Experimental Subgroup A topics experienced improvement in performance ranging from 41% – 53% (Cohen's $d = 0.82 - 1.10$), while one topic experienced 42% decline. Two out of three Experimental Subgroup B topics experienced improvement in performance ranging from 8% – 35% (Cohen's $d = 0.17 - 1.22$), while one topic experienced 12% decline. Eight out of nine Control group (C) topics experienced improvement in performance ranging from 5% – 47% (Cohen's $d = 0.10 - 0.97$), while one topic experienced 11% decline.

Table IV-1. Short-term retention (average in-lab item difficulty vs. end-of-block exam item difficulty) of radiologic anatomy topics.

A = Experimental Subgroup A (topics repeatedly tested via FR items;

B = Experimental Subgroup B (topics repeatedly tested via MCQs);

C = Control group (topics tested only once, via MCQs);

Cohen's *d*-Effect size (reported only for the gains)

	Item difficulty in in-lab testing Session 1	Item difficulty in in-lab testing Session 2	Item difficulty in in-lab testing Session 3	Average item difficulty (standard deviation) across Sessions 1 – 3	Item difficulty (standard deviation) in end of curricular block exam	Gain or loss on end of curricular block exam, compared with average item difficulty across sessions 1–3
A1	0.80	0.83	1.00	0.88 (0.32)	0.46 (0.50)	42% decline (Cohen's <i>d</i> = -1.00)
A2	0.67	0.73	0.01	0.47 (0.50)	0.88 (0.32)	41% gain (Cohen's <i>d</i> = 0.82)
A3	0.92	0.27	0.27	0.48 (0.50)	0.97 (0.17)	49% gain (Cohen's <i>d</i> = 0.98)
A4	0.17	0.36	0.59	0.37 (0.48)	0.90 (0.30)	53% gain (Cohen's <i>d</i> = 1.10)
B1	0.42	0.48	0.49	0.47 (0.50)	0.82 (0.38)	35% gain (Cohen's <i>d</i> = 0.70)
B2	0.76	0.78	0.63	0.72 (0.45)	0.80 (0.40)	8% gain (Cohen's <i>d</i> = 0.17)
B3	0.19	0.57	0.57	0.44 (0.49)	0.32 (0.46)	12% decline (Cohen's <i>d</i> = -0.25)
C1	0.91 (0.28)			-	0.80 (0.40)	11% decline (Cohen's <i>d</i> = -0.31)
C2	0.56 (0.49)			-	0.88 (0.32)	32% gain (Cohen's <i>d</i> = 0.65)
C3		0.44 (0.49)		-	0.91 (0.28)	47% gain (Cohen's <i>d</i> = 0.95)
C4			0.80 (0.40)	-	0.95 (0.21)	15% gain (Cohen's <i>d</i> = 0.37)
C5			0.26 (0.44)	-	0.69 (0.46)	43% gain (Cohen's <i>d</i> = 0.97)
C6	0.67 (0.47)			-	0.80 (0.40)	13% gain (Cohen's <i>d</i> = 0.27)
C7			0.37 (0.48)	-	0.61 (0.49)	24% gain (Cohen's <i>d</i> = 0.50)
C8		0.66 (0.48)		-	0.71 (0.45)	5% gain (Cohen's <i>d</i> = 0.10)
C9			0.38 (0.48)	-	0.64 (0.48)	26% gain (Cohen's <i>d</i> = 0.54)

Table IV-2 displays short-term retention of non-radiologic anatomy topics. All three Experimental Subgroup A topics experienced improvement in performance ranging from 51% – 73% (Cohen's *d* = 1.02 – 1.70). On the other hand, only two out of five Experimental Subgroup B topics experienced improvement in performance ranging from 28% to 55% (Cohen's *d* = 0.57 – 1.22), while the other three topics experienced declines ranging from 2 – 41%. All four Control group (C) topics experienced improvement in performance ranging from 4% – 58% (Cohen's *d* = 0.08 – 1.23).

Long-term Retention

Table IV-3 displays long-term retention (end-of-block exam item difficulty vs. end-of-year quiz item difficulty) of the nine blindly selected radiologic anatomy topics.

Table IV-2. Short-term retention (average in-lab item difficulty vs. end-of-block exam item difficulty) of non-radiologic anatomy topics.
A = Experimental Subgroup A (topics repeatedly tested via FR items);
B = Experimental Subgroup B (topics repeatedly tested via MCQs);
C = Control group (topics tested only once, via MCQs);
Cohen's d – Effect size (reported only for the gains).

	Item difficulty in in-lab testing Session 1	Item difficulty in in-lab testing Session 2	Item difficulty in in-lab testing Session 3	Average item difficulty (standard deviation) across Sessions 1 – 3	Item difficulty (standard deviation) in end of curricular block exam	Gain or loss on end of curricular block exam (compared with average item difficulty across sessions 1–3)
A5	0.59	0.44	0.31	0.45 (0.50)	0.96 (0.19)	51% gain (Cohen's d = 1.02)
A6	0.02	0.20	0.52	0.24 (0.43)	0.97 (0.17)	73% gain (Cohen's d = 1.70)
A7	0.29	0.38	0.45	0.37 (0.48)	0.97 (0.17)	60% gain (Cohen's d = 1.23)
B4	0.36	0.03	0.50	0.29 (0.45)	0.84 (0.36)	55% gain (Cohen's d = 1.22)
B5	0.84	0.53	0.66	0.68 (0.47)	0.66 (0.48)	2% decline (Cohen's d = -0.04)
B6	0.68	0.69	0.51	0.63 (0.49)	0.91 (0.28)	28% gain (Cohen's d = 0.57)
B7	0.65	0.53	0.42	0.53 (0.50)	0.12 (0.33)	41% decline (Cohen's d = -0.96)
B8	0.44	0.67	0.66	0.59 (0.49)	0.56 (0.49)	3% decline (Cohen's d = -0.06)
C10	0.81 (0.39)			-	0.87 (0.37)	6% gain (Cohen's d = 0.15)
C11			0.61 (0.49)	-	0.72 (0.45)	11% gain (Cohen's d = 0.22)
C12		0.33 (0.47)		-	0.91 (0.28)	58% gain (Cohen's d = 1.23)
C13	0.56 (0.49)			-	0.60 (0.49)	4% gain (Cohen's d = 0.08)

Table IV-3. Long-term retention (end-of-block exam vs. end-of-year quiz item difficulty) of radiologic anatomy topics.

A = Experimental Subgroup A (topics repeatedly tested via FR items);

B = Experimental Subgroup B (topics repeatedly tested via MCQs);

C = Control group (topics tested only once, via MCQs);

Cohn's d = Effect size reported only for the gains).

	Item difficulty (standard deviation) in end of curricular block exam	Item difficulty (standard deviation) in end of the year quiz	Gain or loss on end of the year quiz, compared with end of curricular block exam
A1	0.46 (0.50)	0.35 (0.48)	11% decline (Cohen's $d = -0.22$)
A2	0.88 (0.32)	0.50 (0.50)	38% decline (Cohen's $d = -0.90$)
A3	0.97 (0.17)	0.71 (0.45)	26% decline (Cohen's $d = -0.76$)
B1	0.82 (0.38)	0.44 (0.49)	38% decline (Cohen's $d = -0.86$)
B2	0.80 (0.40)	0.77 (0.42)	3% decline (Cohen's $d = -0.07$)
C1	0.80 (0.40)	0.83 (0.38)	3% gain (Cohen's $d = 0.07$)
C2	0.88 (0.32)	0.92 (0.27)	4% gain (Cohen's $d = 0.12$)
C7	0.61 (0.49)	0.63 (0.49)	2% gain (Cohen's $d = 0.04$)

All three Experimental Subgroup A topics experienced declines ranging from 11% to 38%. Similarly, both Experimental Subgroup B topics experienced declines ranging from 3% to 38%. However, all three Control group (C) topics experienced gains in retention ranging from 2% – 4% (Cohen's $d = 0.07 - 0.12$).

Table IV-4 displays long-term retention of the five blindly selected non-radiologic anatomy topics. Both Experimental Subgroup A topics experienced declines ranging from 10% to 49%. On the other hand, two out of three Experimental Subgroup B topics experienced declines in retention ranging from 5% to 10%, while performance on one topic exhibited 48% gain in long-term retention (Cohen's $d = 1.45$).

Table IV-4. Long-term retention (end-of-block exam item difficulty vs. end-of-year quiz item difficulty) of non-radiologic anatomy topics.

A = Experimental Subgroup A (topics repeatedly tested via FR items);

B = Experimental Subgroup B (topics repeatedly tested via MCQs);

C = Control group (topics tested only once, via MCQs);

Cohn's d = Effect size reported only for the gains).

	Item difficulty (standard deviation) in end of curricular block exam	Item difficulty (standard deviation) in end of the year quiz	Gain or loss on end of the year quiz, compared with end of curricular block exam
A5	0.96 (0.19)	0.47 (0.50)	49% decline (Cohen's $d = -1.45$)
A7	0.97 (0.17)	0.87 (0.37)	10% decline (Cohen's $d = -0.28$)
B4	0.84 (0.36)	0.79 (0.41)	5% decline (Cohen's $d = -0.10$)
B6	0.91 (0.28)	0.81 (0.40)	10% decline (Cohen's $d = -0.28$)
B7	0.12 (0.33)	0.60 (0.49)	48% gain (Cohen's $d = 1.45$)

Discussion

This was a study on the effect of testing on retention of knowledge of human anatomy in the context of Year 1 medical education. The study was conducted over the span of one academic year at a medical school with a hybrid Patient-Centered Learning curriculum. Here are few observations based on obtained results.

Firstly, collective performance may not experience a steady improvement when *different* questions are used in repeated testing of a topic. Tables IV-1 and IV-2 highlight this finding. Only three (A1, A4, B2), out of the total seven repeatedly tested radiologic anatomy topics (A1 – B3) experienced steady improvement in performance across the three in-lab testing sessions (Table IV-1). Similarly, only two (A6 and A7), out of the total eight repeatedly tested non-radiologic anatomy topics (A5 – B8) experienced steady improvement in performance across the three in-lab testing sessions (Table IV-2).

Conversely, majority of the repeatedly tested radiologic and non-radiologic anatomy topics (nine out of the total fifteen) experienced fluctuation in performance across the three in-lab tests (Tables IV-1 and IV-2). Performance on some topics showed improvement in the second in-lab testing session, but declined in the third testing session (e.g., topics A2 and B3). On the other hand, performance on a few other repeatedly tested topics showed decline in the second in-lab testing session, but improved in the third session (e.g., topics B1 and B5). This fluctuation in performance on various topics is not an unexpected observation. A similar fluctuation was noted in the study published by Larsen et al. on the effect of repeated testing on final recall of two topics among postgraduate medical trainees³. However, there are subtle differences between the method used by Larsen et al. and the one used in the study presented here. The interval

between each repeated test (total three tests) in the Larsen et al. study was 2 weeks, while the interval between each repeated test (total three tests) in our study was one week. Moreover, Larsen et al. used *same* questions in repeated testing of each topic, while the study presented here used *different* questions in repeated testing of each topic (examples: Figure IV-1). Therefore, the fluctuation in performance on repeatedly tested topics in our study may either be stemming from the smaller (1 week) interval between successive tests, or the usage of *different* questions with different levels of inherent difficulty.

Secondly, topics tested thrice via free-response questions tend to be recalled, four weeks later, to a slightly greater extent than the topics tested thrice, or once, via multiple-choice questions. Tables IV-1 and IV-2 highlight this finding. Radiologic anatomy topics tested thrice via free-response questions (A1 – A4) exhibited gains in retention ranging from 41% – 55% (Table IV-1). On the other hand, radiologic anatomy topics tested thrice via multiple-choice questions (B1 – B3) exhibited gains in retention ranging from 8% – 35%, and radiologic anatomy topics tested once via multiple-choice questions (C1 – C9) exhibited gains in retention ranging from 5% – 47%. The slight superiority of the level of gain from repeated testing via free-response questions was true for non-radiologic anatomy topics as well (Table IV-2). Non-radiologic anatomy topics tested thrice via free-response questions (A5 – A7) exhibited gains in retention ranging from 51% – 73%. On the other hand, among the five non-radiologic anatomy topics tested thrice via multiple-choice questions (B4 – B8), only two topics (B4 and B6) exhibited gain in short-term retention; the extent of gain for topic B4 was 55% and for topic B6 was 28%. Similarly, non-radiologic anatomy topics tested once via multiple-choice questions (C10 – C13) exhibited gains in retention ranging from 6% – 58%. Based on

these findings, one can surmise that retrieval practice through free-response questions tends to enhance retention of knowledge to a generally greater extent indeed than retrieval practice through multiple-choice questions. However, one must note that the magnitude of difference among gains through retrieval practice by either means is not large. The generally higher retention of topics repeatedly tested by free-response questions shows that retrieval practice via this method may be marginally more useful in fostering retention of content for up to four weeks, than retrieval practice via multiple-choice questions. This finding is in chorus with previous discussions and reports on the added benefit of repeated retrieval practice via free-response questions.^{2,3,7-9} However, medical educator scholars must be wary of the current tradition of in-house assessment in pre-clinical medical education. In-house assessment in basic medical sciences in general, and human anatomy in particular, relies heavily on multiple-choice questions. And, repeated usage of free-response questions in frequent in-lab testing may be confusing for students in regards to the nature of final high-stakes exam for that content. Therefore, to any anatomy educators contemplating using frequent testing via free-response questions as a learning tool, we recommend clarifying the purpose of such testing to their students, i.e. slightly enhanced retention of knowledge for a up to a month's duration.

Thirdly, repeated testing with either method (free-response or multiple-choice questions) may not be of much help in long-term knowledge retention. Tables IV-3 and IV-4 highlight this finding. All radiologic and non-radiologic anatomy topics tested thrice via free-response questions (A1, A2, A3, A5 and A7) exhibited a decline in performance ranging from 10% to 49% between end-of-curricular-block exams and end-of-the-year quiz. Similarly, four out of five topics tested thrice via multiple-choice

questions (B1, B2, B4 and B6) exhibited decline in performance ranging from 3% to 38%, while only one topic (B7) exhibited a 48% gain. Moreover, the length of interval between end-of-year quiz and end-of-block exams was found to be unrelated to the strength of declines in retention. For example, topic A1, that was tested seven months prior to the end-of-year quiz, experienced smaller decline in retention than topic A3 that was tested only two months prior to the end-of-year quiz (11% vs. 26%) (Table IV-3). The decline in retention of knowledge of various topics despite repeated testing thereof is contrary to earlier reports.^{5,8,9} The studies published by Larsen et al.⁵, Butler and Roediger,⁸ and McDaniel et al.,⁹ reported enhanced retention of knowledge over long-term durations via repeated testing of content. However, a closer look would reveal that the definition of “long-term” in some of those studies differed from the one used in our study. For example, Butler and Roediger used one-month as the definition of long-term duration.⁸ Similarly, McDaniel et al. used a range of 30 to 56 days (average 40 days) as criteria for long-term interval.⁹ On the other hand, our study considered 2 to 7 months as long-term duration. It is worth noting that the studies by Butler and Roediger and McDaniel et al. were conducted in the context of cognitive psychology and reported in a journal relevant to that field. However, our study was conducted in the context of Year 1 medical education in a real-life educational setting. What might be considered as long-term interval between exposure (repeated testing) and outcome (cumulative or final exam) in experiments conducted in one domain (cognitive psychology) may not apply to another domain (pre-clinical medical education). Therefore, decline in long-term (2 – 7 months) retention of knowledge of the repeatedly tested topics in our study may not be an unusual finding. However, Larsen et al. intentionally used a long-term interval more akin

to real-life educational setting (6 months) in their study, and found significantly enhanced retention courtesy of repeated testing of content.⁵ One must note that the Larsen et al. study was conducted in the context of postgraduate medical education, in which trainees (medical residents) are required to participate in at least one or two didactic conferences per year as part of their training program. On the other hand, the context of our study was Year 1 medical education; a curriculum that is packed with a variety of teaching and learning experiences in a variety of basic medical science subjects. Therefore, the contrary finding in our study hints that repeated testing may not be as efficacious for long-term retention in the setting of an already crowded pre-clinical medical curriculum.

Fourthly, testing only once with multiple-choice questions may be enough for long-term (2 – 7 months) knowledge retention of Year 1 radiologic anatomy content. Table IV-3 highlights this finding. Radiologic anatomy topics tested once via multiple-choice questions (C1, C2 and C7) experienced gains in retention ranging from 2% - 4% between end-of-curricular-block exams and end-of-the-year quiz. Here too, the degree of retention was not related to the interval between end-of-year quiz and end-of-block exam. For example, topic C1 that was tested seven months prior to the end-of-year quiz, exhibited a slightly greater gain in retention than topic C7 that was tested just two months prior to the end-of-year quiz (3% vs. 2%) (Table IV-3). Two points are worth noting here. One is that these three topics (C1, C2 and C7) were blindly selected for testing in the end-of-the-year quiz since, for practical reasons, we could not test all experimental and control group topics in the end of the year quiz; the quiz was kept deliberately short in order to encourage voluntary participation from the students. The other point is that, although the long-term gain in retention exhibited in the knowledge of these topics is

numerically small (2% - 4%), it is still of considerable value owing to the numerically greater declines in retention of all the topics repeatedly tested via free-response (10% to 49% declines) and four out of five topics repeatedly tested via multiple-choice (3% to 38% declines) questions. This finding raises an interesting question. Previous studies have reported the superiority of *repeated* testing over *once* testing of content.^{1,2,13} We wonder why this mnemonic benefit of test-enhanced learning could not be reproduced in our study. Could it be that the memory imprint of radiographic images shown and discussed only once lasts longer than the memory imprint of different images shown to repeatedly emphasize the same concept? Or was our finding an artifact of subject or topic characteristics? Expansion of this study will help find a definite answer in this regard.

A few limitations apply to the findings in this study. First is whether three *different* questions on the same topic classify as *repeated* testing at all. The study by Logan et al. makes a strong case in this regard.¹¹ In our study, questions to repeatedly test the same topic were constructed intentionally and carefully (examples: Figure IV-1). Moreover, the interactive discussion following every in-lab testing session was guided to emphasize (and re-emphasize) the same topic. However, unconsciously, we might still have tested different concepts in repeated tests of the same topic. If this happened, we might have helped create different memory imprints for each *question*, rather than help solidify the imprint of the same *memory*, thereby confounding our result. As Butler and Roediger⁸ and McDaniel et al.⁹ put it, testing enhances the storage strength (relative permanence) as well as the retrieval strength (momentary accessibility) of memory traces and repeated retrieval practice is meant to enhance these strengths for the purpose of

deeper and longer-lasting memory imprints. To what degree, in our study, did repeated testing via different questions help in enhancing a concept's storage and retrieval strengths is open to interpretation. Perhaps a more controlled design, such as repeated testing of the same topic via free-response questions in a random half of subjects and via multiple-choice questions in the other random half of subjects, will help derive more sound conclusions in this regard.

The second limitation is that although the setting of our experiment was gross anatomy laboratory, selective study (or lack thereof) of some topics might have occurred outside the laboratory setting that was beyond the control of the investigators. Such selective study might have influenced the outcome in any or all groups of topics (experimental and control) used in our study. We have now become cognizant of the need to either control, or document, the exposure to the under-investigation topics outside the anatomy laboratory setting (such as Patient-Centered Learning discussions, clinical skills sessions etc.) in order to derive more definite conclusions from our findings. The study by Larsen et al.⁵ provides an excellent example of how such documentation helps in strengthening the relationship between the outcome (e.g., knowledge retention) and the intervention (e.g., repeated testing).

The third limitation is the size and scope of our study. We used a total of 28 (17 radiologic and 11 non-radiologic anatomy) topics in our study. This study may be useful as a pilot. In order to draw more definitive and generalizable conclusions, the number of topics in each group (experimental and control) should be increased and a variety of topics from anatomical (histology, embryology, gross- and neuro-anatomy) as well as other basic medical sciences (physiology, biochemistry, pharmacology, etc.)

should be included in the study. While we plan on expanding our study, we invite scholars located elsewhere to conduct studies on the role of test-enhanced learning in the context of pre-clinical medical education.

By and large, as educators of anatomy, we have found no stakes testing to be a beneficial educational intervention in line with current emphases on using assessment to drive and promote learning.^{14, 15} Awareness among medical educators as to what content matters and how best to reinforce its learning is, first and foremost, beneficial for the learners. Moreover, such an awareness of the relevant content and mental operations behind its learning brings an evidence-base to their teaching and learning practices that eventually benefits the overall medical education system. Findings of our study suggest that educators of human anatomy who want their students to retain the acquired knowledge, so it could be applied to clinical situations more expeditiously, might consider using no-stakes production tests (such as free-response type questions) for ongoing self-assessment and feedback on curricular content. With an appropriate setting and provision of feedback, the direct and indirect effect of such tests may help raise competence of future physicians and health science professionals.

References

1. Karpicke JD, Roediger HL III. The critical importance of retrieval for learning. *Science*. 2008;319:966–968.
2. Roediger HL III, Karpicke JD. The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci*. 2006;1:181–210.
3. Larsen DP, Butler AC, Roediger HL. Test-enhanced learning in medical education. *Med Educ*. 2008;42:959–966.
4. Kromann CB, Jensen ML, Ringsted C. The effect of testing on skills learning. *Med Educ*. 2009;43:21–27.
5. Larsen DP, Butler AC, Roediger HL. Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. *Med Educ*. 2009;43:1174–1181.
6. Dobson J. Retrieval practice is an efficient method of enhancing the retention of anatomy and physiology information. *Adv Physiol Educ*. 2013;37:184–191.
7. McDaniel MA, Roediger HL III, McDermott KB. Generalizing test-enhanced learning from the laboratory to the classroom. *Psychon Bull Rev*. 2007;14:200–206.
8. Butler AC, Roediger HL. Testing Improves Long-Term Retention in a Simulated Classroom Setting. *Eur J Cogn Psychol*. 2007;19:514–527.
9. McDaniel MA, Anderson JL, Derbish MH, Morrisette N. Testing the testing effect in the classroom. *Eur J Cogn Psychol*. 2007;19:494–513.
10. Foos PW, Fisher RP. Using tests as learning opportunities. *J Educ Psychol*. 1988;80:179–83.

11. Logan JM, Thompson AJ, Marshak DW. Testing to enhance retention in human anatomy. *Anat Sci Educ*. 2011;4(5):243–248.
12. Hojat M, Xu G. A visitor’s guide to effect sizes—statistical significance versus practical (clinical) importance of research findings. *Adv Health Sci Educ*. 2004;9(3):241–249.
13. Butler, A.C. Repeated testing produces superior transfer of learning relative to repeated studying. *J Exp Psychol Learn Mem. Cogn*. 2010;36,1118–1133.
14. Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: An unavoidable truth? *Anat Sci Educ*. 2009;2:199–204.
15. Wood T. Assessment not only drives learning, it may also help learning. *Med Educ*. 2009;43:5–6.

CHAPTER V

SUMMARY AND CONCLUSION

The dissertation presented here is a composite of studies on two major components of a curriculum: “Educational Strategies” and “Assessment”. Chapter I (Introduction) briefly describes these two components. Chapters II and III describe two separate investigations on the quality of assessment, while Chapter IV elaborates on the usefulness of retrieval practice as a learning strategy in pre-clinical medical education. Here are a few conclusions based on the major findings presented in this dissertation.

Examinee performance on the free-response and multiple-choice versions of an exam may consistently differ from each other (Chapter II). The consistently disparate performance stems from the version itself; there is guessing and cueing involved in answering multiple-choice questions, and such guessing and cueing may be minimal in answering the free-response version of the same items.¹ In other words, performance on multiple-choice questions may be influenced by factors other than true knowledge of examinees. The weight of that influence may depend on the presence of flaws and implausible distractors (i.e., incorrect options with low attractiveness to the examinees) in multiple-choice questions.

Difficulty of a multiple-choice exam may be lower than expected when distractor functioning in the exam items is low (Chapter II). Most multiple-choice distractors in an exam should be selected by $\geq 5\%$ examinees.² These distractors should be selected by examinees that use partial knowledge to answer a question. Partial knowledge allows

examinees to rule-in or rule-out various multiple-choice distractors based on the plausibility of those distractors. A distractor's selection frequency is a representation of its plausibility, hence attractiveness, to the examinees. An item writers' ability to accurately gauge examinee knowledge is adversely affected by low attractiveness, hence "functioning", of multiple-choice distractors.^{1,2} This inadvertently allows test-wise (and not necessarily well-prepared) students to perform well on an exam, which is an unwanted outcome of high-stakes assessment in undergraduate medical education.

When examinees are provided a free-response (i.e., short-answer) version of an item, their incorrect responses may be used to construct more plausible multiple-choice distractors. Incorporation of those distractors into the multiple-choice questions may help reduce the disparity between expected and observed difficulty of a multiple-choice exam. In other words, such a maneuver allows more apt assessment of examinee knowledge thereby allowing the examiner to make valid (accurate) and reliable (reproducible) conclusions from scores obtained on the exams. Validity and reliability have specific meanings in terms of educational assessment.³ Validity refers to "true knowledge representativeness" of scores obtained on an exam and can be strengthened through collection of evidence from various sources. Two such sources are "relations to other variables" and "internal structure", which were discussed in detail in Chapters II and III of this dissertation.

Increased distractor functioning may increase the reliability of scores obtained on multiple-choice exams (Chapter II). The concept of reliability is connected to the concept of validity and refers to the reproducibility or consistency of scores from one assessment to another. When new distractors developed from incorrect responses on free-

response version of the items are incorporated in an exam's items, their increased selection helps elicits a greater range of abilities from the examinees. Greater ability range manifests itself in the form of increased standard deviation, and it is the standard deviation that has a directly proportional relationship with the reliability coefficient of scores obtained on a multiple-choice exam. It is worth noting, however, that the reliability coefficient may be higher, despite low quality of an exam, if more unprepared (or low ability) examinees are encouraged to take an exam. Low performance of some unprepared (or low ability) examinees may artificially increase the standard deviation of scores, despite low quality of the exam owing to low distractor functioning. Therefore, reliability coefficient should be interpreted cautiously with appropriate attention to overall quality of the exam vis-à-vis distractor functioning and discriminatory ability of the items used in an exam.

Removal of item flaws (along with enhancement of cognitive level tested by the items), as well as replacement or removal of non-functioning (less than 5% selection frequency) distractors may increase distractor functioning in multiple-choice questions (Chapter III). Increased distractor functioning resulting from such interventions may help in improving the discriminatory ability of multiple-choice questions, thereby allowing better separation of high-performing and low-performing students. Our findings suggest equal usefulness of both these interventions in raising discriminatory ability of multiple-choice questions used in assessment of knowledge of the basic medical sciences. With enhancement of discriminatory ability, performance on an item can be attributed more clearly to knowledge of the topic under assessment, with less influence of any extraneous factors such as guesswork or confusing stem or options. Such extraneous factors account

for some variance in performance on multiple-choice exams. Therefore, enhanced discriminatory ability of multiple-choice questions may prevent performance variance introduced by extraneous factors from weakening the evidence of validity of obtained scores.³

Knowledge of human anatomy topics repeatedly tested via free-response questions is retained to a greater extent than knowledge of topics repeatedly or once tested via multiple-choice questions (Chapter IV). This finding applies to topics in both radiologic and non-radiologic (clinical) anatomy. In our study, knowledge of topics tested thrice via free-response questions was retained, four weeks later, to a greater extent than knowledge of topics tested thrice or once via multiple-choice questions. This finding is in unison with previously published reports, and attests to the greater usefulness of repeated production tests (such as free-response questions) over repeated or singular recognition tests (such as multiple-choice questions) in short-term (up to four weeks) retention of knowledge.⁴

Knowledge of radiologic anatomy topics tested once via multiple-choice questions may be retained to a greater extent than knowledge of radiologic and non-radiologic anatomy topics tested repeatedly via free-response or multiple-choice questions (Chapter IV). In our study, end-of-the-year performance on once-tested control group of topics showed an unexpected gain in knowledge retention of a small magnitude. On the other hand, end-of-the-year performance on all but one thrice-tested experimental group of topics showed small to moderate declines in retention. The higher level of long-term retention of once-tested radiologic anatomy content raises some interesting questions and the study warrants expansion to explore this phenomenon in further detail.

The work presented in this dissertation shows that the quality improvement of teaching and assessment in medical education is a scholarly process. This process entails clear definition of learning goals and objectives, adequate integration of educational content with teaching and learning strategies, and usage of valid methods to assess the knowledge acquired through a curriculum. When this process follows a systematic approach based on theories of assessment and learning, it benefits not only the learner but the instructor as well, by providing a venue for educator scholarship as well as professional and personal growth.⁵ As is the case across the US and Canada, most basic science faculty members are experts in their field, but may not be familiar with the jargon, methods and resources used in quality improvement of undergraduate curricula. Owing to their heavy involvement in pre-clinical education of future physicians, basic science faculty should have at least some elementary skills in regards to evaluating their own teaching and assessment practices. Such skills will help reform and redesign pre-clinical curricula to fulfill the requirements of rapidly advancing basic medical sciences and conceptualization of “competence” amongst future physicians⁶.

The field of medical education requires constant revision of teaching formats, curricular content, and assessment techniques, and the dissertation work presented here keeps up with the contemporary traditions in this area. This work is in sync with the mantra that assessment drives learning, and that high quality assessment is one of the tenants of competency-based medical education. What one hopes to achieve through such investigation is to allow only the competent students to progress to the next stage of training and help inform a sound remediation process for the rest. Through constant contemplation on assessment and learning practices, we enrich the culture of research in

medical education with ultimate benefit for the medical profession and society as a whole.

References

1. Haladyna TM, Downing SM. How many options is enough for a multiple-choice test item? *Educ Psychol Meas.* 1993;53:999–1009.
2. Rodriguez, MC. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educ Measure Issues Prac.* 2005;24(2);3–13.
3. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am J Med.* 2006;119(166):67–116.
4. Larsen DP, Butler AC, Roediger HL. Test-enhanced learning in medical education. *Med Educ.* 2008;42:959–966.
5. Glassick CE, Huber MT, Maeroff GI. *Scholarship Assessed: Evaluation of the Professoriate.* San Francisco (CA): Josey-Bass Publishers; 1996.
6. Parsal GJ, Bligh J. The changing context of undergraduate medical education. *Postgrad Med J.* 1995;71(837):397–403.