AN ANALYSIS OF TEST CONSTRUCTION PROCEDURES AND
SCORE DEPENDABILITY OF A PARAMEDIC
RECERTIFICATION EXAM

by

Ingrid Maria de Vries

A thesis submitted to the Faculty of Education
in conformity with the requirements for
the Degree of Master of Education

Queen’s University
Kingston, Ontario, Canada
September, 2012

Copyright ©Ingrid de Vries, 2012
Abstract

High-stakes testing is used for the purposes of providing results that have important consequences such as certifications, licensing, or credentialing. The purpose of this study was to examine aspects of an exam recently written by flight paramedics for recertification and make recommendations for development of future exams. In 2008, an unexpectedly high failure led to revisions in the exam development process for flight paramedics. Using principles of classical test theory and generalizability theory, I examined the decision consistency and dependability of the examination and found the decision consistency for dichotomous items to be within acceptable limits, yet the dependability was low. Discrimination was strong at the cut-score. An in-depth look into the process used to set the exam, as well as the psychometric properties of the exam and the items have led to recommendations that will contribute to future development of dependable exams in the industry that result in more valid interpretations with respect to paramedic competence.
Acknowledgements

Completing a graduate degree has been my goal for many decades, and I am fortunate to have been able to go through this journey of learning. None of this would have been possible without the continued support and encouragement of the people around me.

To Dr. Denise Stockley, my first advisor and committee member, I thank you for suggesting I take a course in measurement theory; this sparked my interest in measurement and evaluation, and guided me through all of graduate school. To Dr. Don Klinger, my thesis advisor and academic mentor, thank you for your patience and support; mention should be made that I was almost scared out of graduate school on the first day of class when I heard that the course in measurement theory was “challenging and rigorous”.

None of this would have been possible, nor have been initiated, without the influence of my family. My parents, John and Gerda, instilled the value of higher education from the time my father was pursued a PhD and continuing until my mother fulfilled her life-long dream to complete a university degree after raising a family.

To my children, Maria, Nick, and Damian, I hope that you have learned that learning never ends; your years of support, patience, and humour were helpful in getting me through to the end. Finally, a huge thank you to my husband, Michael, for tolerating my many absences from family life and for picking up the household slack so that the dream could become reality.
# Table of Contents

Abstract .................................................................................................................................................. ii
Acknowledgements .................................................................................................................................. iii
List of Figures .......................................................................................................................................... v
List of Tables ........................................................................................................................................... vi
Chapter 1 Introduction .......................................................................................................................... 1
Chapter 2 Literature Review .................................................................................................................. 13
Chapter 3 Methodology ....................................................................................................................... 32
Chapter 4 Results .................................................................................................................................. 38
Chapter 5 Discussion ............................................................................................................................ 51
References ............................................................................................................................................. 69
Appendix A Letter of Information ......................................................................................................... 77
Appendix B Consent Form ...................................................................................................................... 79
Appendix C Exam Objectives ................................................................................................................ 81
Appendix D Interview Questions .......................................................................................................... 84
Appendix E Analysis of Dichotomous Items ....................................................................................... 85
Appendix F Analysis of Discriminating Items ..................................................................................... 88
Appendix G Analysis of Problematic Items ......................................................................................... 90
List of Figures

Figure 1 Frequency Distribution for Exam Scores ................................................................. 40
Figure 2 Correlation between Predicted and Actual p-values for All Items ................................. 41
Figure 3 Correlation between Predicted and Actual p-values for Multiple Choice Items ............ 42
Figure 4 Correlation between Predicted and Actual p-values for Constructed Response Items.... 42
List of Tables

Table 1 Summary of Descriptive Statistics for Recertification Exams for 2008 and 2010 .......... 39
Table 2 Estimated G- Study Variance Components from 2010 Recertification Exam.................. 46
Table 3 Decision Study Results ................................................................................................. 47
Chapter 1

Introduction

High-stakes testing is used for the purposes of providing results that have important consequences for those writing the tests. These tests include those used for certifications, licensing, or credentialing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). These examinations or tests, terms which are used interchangeably in this study, are commonly used for initial and ongoing certification in many professions. I became interested in high-stakes testing for paramedics after there was an unusually high failure rate in a flight paramedics’ written recertification exam conducted in 2008 in Ontario. In previous years, very few paramedics failed the exam, and the employer and delegating physician group believed that the knowledge level of the practicing paramedics was high enough to grant certification to all the paramedics, granting certification despite the failing scores. The results from the 2008 administration created a challenge for granting certification to all of the paramedics given the much higher failure rate. As a result, it was decided to revise the examination process for subsequent administrations.

In general, changes in pass-fail rates from one administration to another can be explained by varying abilities of different examinee cohorts or changes in knowledge levels of same cohorts. However, in the case of the flight paramedics, most of the
examinees in 2006 were the same people as those in 2008. There had been no clear change in paramedic demographics, training programs, or ongoing training to explain the increased number of failures.

The purpose of the test and the type of measurement dictate which format of testing is most appropriate. Tests can be *norm-referenced* where examinee scores are ranked relative to the other examinees, or *criterion-referenced* where an examinee score is related to a standard such as a cut-score. In norm-referenced testing, candidates receive a score that lies along a continuum of scores. In criterion-referenced testing, candidates with a measured score exceeding a predefined passing grade, also known as a cut-score, are classified as *masters*. Those with a score below the cut-score are *non-masters* (Crocker & Algina, 1986, pg. 411).

The flight paramedic exam is considered to be criterion-referenced and paramedics have been classified as certified or not certified according to their raw scores. An arbitrary cut-score of 70% was applied to all of the exams that have been administered. My review of the exam procedures revealed that there was no development process in place, other than a collation of questions that were created by several physicians. No systemic process was followed to guide the exam development and no analyses were conducted on the items or the exam.

As Hambleton and Slater (1997) described, significant progress has been made in exam development and the psychometric characteristics of credentialing exams, especially around the areas of reliability and decision consistency, defined as “the

Despite the wealth of research with respect to certification and credentialing examinations, little if any research has been done to examine the credentialing exams for paramedics, and these examinations have important consequences for doctors, paramedics, and their patients. If a paramedic is incorrectly classified as competent according to her/his observed exam score, and is allowed to continue to practice, there is potential for the paramedic to deliver incorrect patient treatment and poor patient outcomes. On the other hand, if a paramedic is classified as not competent, and truly is competent, her/his certification could be removed, resulting in unfair dismissal from the employer and loss of income for the paramedic.

The consequences of the paramedic recertification examination support the importance of administering a sound criterion-referenced exam that accurately separates competent from non-competent paramedics. This means that all attempts should be made
to minimize the sources of measurement error. The stakeholders in the examination process have a vested interest in knowing that the instrument (the exam) enables valid interpretations to be made and that a defensible approach has been followed in its development. In the case of the paramedic recertification examination, the main stakeholders in the exam are the group of physicians who delegate medical acts to the paramedics, and the paramedics who write these exams. Medico-legal implications, such as litigations, are directly associated with a physician’s license to practice. The paramedics look for assurance that the exam development process is fair, and reflective of the expected knowledge and scope of practice.

Purpose

In the interest of increasing patient safety and reducing medico-legal risk to the group of delegating physicians, as a result of making correct classification decisions on the paramedic recertification examination, I studied the examination data collected during the 2010 year. This appears to be the first time a systematic development process was used in high-stakes testing for paramedics in Ontario. What has not yet been examined is how the changes in process have affected the passing rate, and what further modifications can be made to improve future exams. I was particularly interested in the changes in process that were followed as a direct result of the poor outcomes two years earlier in 2008. I investigated the process used to set the exam and conducted psychometric analyses of the exam and the items to guide the following questions of this study:
1. What systemic approaches were used in the 2010 paramedic recertification exam process and how did these changes potentially impact the measured scores?

2. What was the decision consistency for this exam, and what were the factors that impacted this?

3. What recommendations can be made to develop a more consistent dependable exam in the future?

Results of this study may result in changes in policy and practice at the local or provincial levels for the written exams; lessons learned from this study could be applied to the practical examinations that paramedics are expected to complete.

Rationale

Though paramedic certification and recertification exams have been conducted for many years throughout Canada and the United States, a comprehensive framework for this process does not currently exist. Standards, including frequency and format vary according to the employer and regulator. In Ontario, Canada, there are four certification levels: Primary Care, Advanced Care (land), Advanced Care (flight), and Critical Care, the latter two applying to flight paramedics. Training is done at post-secondary institutions such as community colleges, but certification to practice must be issued through a base hospital program since Ontario-based paramedics work under the delegation of a physician’s medical license. Specialized training is required for flight paramedics as they typically transport very ill patients over a long distance. These
paramedics are providing care at the level of intensive care without a physician on board, and hence, their training and maintenance of skills must be thorough. The certification and recertification processes for Ontario’s flight paramedics consist of an annual process of written and practical exams, in alternating years. The practical exam is conducted during odd years in an Objective Structured Clinical Examination (OSCE) style, similar to the assessment process conducted by medical schools around the world.

In my extensive experience as a paramedic in Ontario, I was required to pass the recertification exams in order to maintain my certification. From a paramedic’s point of view, the exam items seemed to measure vague competencies, have poorly worded questions, and poor scores did not appear to have any follow-up. As a paramedic, I questioned what the exams were supposed to measure – competencies that reflected the professional responsibilities or the ability to decipher complicated questions based on seldom used areas of knowledge.

After some time working as a paramedic, I was transferred to a position in the education department for flight paramedics, where I was part of a team responsible, along with a Medical Director, to develop and administer the recertification exams, both written and practical.

The written examinations are administered during the even years. Concerns arose with the unexpectedly high failure rate of 25% across the 82 examinees who wrote the 2008 examination. In comparison, the 2006 examination only had a failure rate of 3.75%. The change in failure rate from 2006 to 2008 led to a review of the existing exam
process. The education department and Medical Director were not able to assess whether the failure rate in 2006 was a true reflection of paramedic performance; we were unable to assess whether either failure rate was too high, too low, or appropriate. Recording keeping and data storage for both 2006 and 2008 was poor, preventing psychometric analysis of the exam results. Triggered by the high failure rate that occurred in 2008, the written exam underwent modifications in the development process for 2010 version. In 2008 and earlier, questions for the written exam were submitted by a number of physicians, with specific construct categories identified beforehand. For example, one physician would have created 10 questions on cardiology, whereas another physician would have been responsible for another 10 questions on the respiratory system. All questions were multiple-choice or short-answer, constructed-response items; multiple-choice items were scored on a two-point scale (right or wrong) and the constructed-response items were worth between one and five marks. The exam items were entered onto a website, and the exam was taken in an unsupervised setting online and graded online. The cut-score for passing was set arbitrarily at 70% as per past practice.

There were no obvious reasons for the drastic change in failure rates between 2006 and 2008. There had been no changes in the method used to create the exam, and the cut-score was always set at 70% based on the raw scores. The paramedics writing the exam were largely the same people who had been writing in previous years. Potential causes of the increased failure rate in 2008 may have been due to a change in the abilities of the personnel writing the exam, changes in the Continuing Medical Education (CME),
improper placement of the cut-scores, the absence of judges to set a cut-score, or problems with the exam items. Was the high failure rate an indication that the exam was more difficult than in the past? Was there a knowledge problem that had surfaced? Were the paramedics’ observed scores reflective of their true scores?

There is no reason to believe that there were any differences in the demographics or skill and knowledge levels amongst paramedics from one year to the next. The initial training programs remained unchanged over the years so newly certified paramedics had met the same requirements to graduate as those before them. The ongoing CME training remained unchanged in content or format, and the CME requirements remained the same: 24 hours of clinical time, 32 hours of didactic delivery, and 24 hours of self-directed education. The doctors judging the scores did not appear to have changed their scoring procedures for the open-ended items. Hence it appears the most likely cause for the differences in pass rates were due to differences in exam difficulty and the impact this had on the arbitrarily set cut score.

In response to the potential flaws that existed in previous exams and the examination process, modifications were made to the 2010 exam, including the use of a cut-score setting process, identified objectives, proctored exam administration, and item editing before publication. The items were referenced to preset objectives that originated from the National Occupational Competency Profile (NOCP), a nationally recognized standard of performance for paramedics of all certifications. Three of the broad categories found in the NOCP Guidelines were the basis for item creation: assessment,
integration, and therapeutics. Within these categories, questions could be allocated to constructs based on body systems: cardiovascular, respiratory, neurology, paediatrics, neonates, obstetrics, gastrointestinal, trauma, immune, and endocrine (see Appendix C). Changes were made to the exam development process in an attempt to standardize the process, reduce the numbers of incorrect classifications, increase consistency, and obtain measured scores that would provide more accurate interpretations of paramedics’ competence (validity).

In 2010, a group of seven physicians created items based on the identified objectives. Submitted questions were screened by members of the education department for accuracy and wording in an attempt to reduce or eliminate flawed items (Downing, 2005), and revised as needed. For example, a long complicated stem was reworded to simplify and shorten the question. The resulting recertification exam consisted of 78 dichotomously scored multiple-choice and constructed-response items. A panel of four judges - two physicians and two Critical Care certified educators - met over three occasions via telephone and in person to determine the cut-score through an Angoff procedure. Written instructions were provided to the judges to explain the process and expectations. Over one iteration of each question, each judge identified the probability of the minimally competent candidate (MCC) getting the correct answer. All probabilities were summed and averaged over the number of items and the judges to arrive at a cut-score of 77%.
The 2010 flight paramedic examination represents the first attempt to use sound exam development procedures to develop the certification/recertification examination to eliminate the issues of previous exams. The revision to the exam process focused on: (a) identification of the objectives, and providing these to the item writers, (b) definition of the type of standard (relative or absolute), (c) use of a panel in a standard setting process to set a cut-score, (d) development of an item construction process, (e) evaluation of exam reliability and validity, (f) administration and analysis of a pilot exam, and (g) documentation of the process from start to finish.

Definition of Terms

Certification is a process conducted by an employer where a candidate demonstrates safe and ethical practice. Licensure, on the other hand, is provided by a professional governing body and gives permission through a license to practice a skill or trade.

Competence is a label associated with performance assessments, especially in a professional environment. To identify competent behavior, the standards for a profession must be clearly defined. If no such standard exists, through an accreditation process, for example, then the performance standards need to be defined. A competent professional is able and capable of satisfactorily performing at this standard.

Cut-score, also known as a cutoff score, is defined as the specific mark or grade that separates a successful examinee from an unsuccessful examinee. Cut-scores are used with both norm-referenced and criterion-referenced examinations. In the case of
criterion-referenced examinations, the determination of the cut-score should be set through a standard setting procedure.

*Flight Paramedics* refers to paramedics who transport patients via airplane, helicopter, or another vehicle, and have specialty training in aero-medical and transport medicine. Certification levels for flight paramedics include Primary Care Flight (PCF), Advanced Care Flight (ACF) or Critical Care (CC). In Ontario, all flight paramedics must hold an Advanced Emergency Medical Care Attendant (A-EMCA) certificate issued by the Ontario Ministry of Health and Long Term Care. After receiving A-EMCA certification, these paramedics proceed through rigorous training programs that allow them to progress through the different levels of certification. Flight paramedics are delegated to perform medical acts through the licensure of transport physicians on staff.

*Objective Structured Clinical Examination (OSCE)* is a procedure used widely in the medical system to evaluate clinical performance of physicians and other allied health care professionals. The format of the exams is comprised of skills stations or scenarios with a mannequin or standardized patient. The results of the OSCE are used wholly or partly to confirm successful completion of a program or to grant certification.

*Performance Assessment* incorporates all the points of assessment that comprise the evaluation of a professional’s performance in his or her role. Performance assessments are based on demonstration of a task, and a grade is allocated to that performance based on a scoring sheet. The task could be a skill, a written evaluation, or a
patient simulated scenario that encompasses patient care, critical thinking, and treatment plans.

_Standard Setting_ is the process of setting a cut-score. The actual process of setting a standard can be lengthy, complex, and costly; the method chosen will depend on various factors. Many methods of standard setting exist including the Angoff, Modified Angoff, Nedelsky, and Bookmark procedures. The methods differ in the manner in which the cut-scores are determined. For example, the Angoff method involves picturing a minimally competent candidate (MCC) to identify the minimal abilities and knowledge levels that are acceptable.

Overview of the Study

This chapter provided an overview of high-stakes testing and the certification process for flight paramedics in Ontario. The purpose of this study is outlined with a detailed rationale for conducting the study. Chapter 2 examines the literature that relates to certification, criterion-referenced testing and relevant methods of analysis, and the standard setting process. Chapter 3 describes the data used in the study and the methodology used for analysis. The results of the analysis are explained in Chapter 4, and finally, Chapter 5 provides a discussion of the findings and recommendations for future exam development.
Chapter 2

Literature Review

Given the nature and significance of the paramedic recertification examinations, and the unexplained variability that has occurred over the recent years, it is important to examine the relevant research and literature. This review of research and literature is based on the key concepts related to criterion-referenced exams and the resulting decisions made based on the measured scores. Investigation into the exam development process is warranted in all cases, but even more so when pass-fail results from an exam are unexpectedly low or high. Many factors can impact the classification results, and efforts should be made to reduce or eliminate sources of potential error. Exploration of these factors will help to focus the discussion, analyses, and recommendations on the areas that potentially impacted the classifications of the paramedics. This literature review provides an exploration of research associated with cut-scores and standard setting, judges, and item development. Decision consistency and exam dependability, important estimates in credentialing exams, along with generalizability theory, are also discussed.

Cut-scores and Standard Setting

Cut-scores represent the score on an examination used to separate examinees into two different levels of performance such as competent/not-competent or pass/fail. Cut-scores can be determined in a variety of ways, most commonly through standard setting
procedures. The process of standard setting has been discussed, researched, and debated for decades with well-known procedures that include the Angoff and modified-Angoff, Bookmark, Contrasting Groups, Examinee Paper Selection, Holistic, and Nedelsky methods (Hambleton, Jaeger, Plake, & Mills, 2000). These methods are centred on either the exam or the examinee, meaning that judgments are based on the perceived item difficulty or anticipated examinee performance. Each standard setting method has advantages and disadvantages that should be considered in the context of the intent and setting. Besides the method chosen, one must consider the number of judges who will set the standard and their qualifications, the number and types of exam items to be reviewed, and the number of iterations to conduct in the standard setting procedure. As Kane, Crooks, and Cohen (1999) stated, it is not the standard setting that is difficult; rather “what is challenging is to choose an appropriate standard setting method and to develop confidence that the resulting passing score is at least approximately at the “right” level (p. 195).

A cut-score can be a relative standard, meaning a predetermined percentage of examinees will pass, or an absolute standard, where a fixed value, or cut-score, represents the cutoff point for pass-fail. Relative standards typically result in a consistent proportion of examinees meeting the standard, yet there may be differences in actual performance across examinations. Muijtjens, Hoogenboom, Verwijnen, and van der Vleuten (1998) describe a relative standard applied to a formative test with Dutch medical students over a 20-year span; using the mean minus standard deviation as a cut point, they found a 15%
failure rate. When they applied the previous relative standard to a new absolute standard, using previous test results (n=27,000), they found that 6% of the pass/fail decisions changed. The conclusion of this study was that the pass/fail rates change considerably depending on which type of standard is being used.

Clauser, Swanson, and Harik (2002) looked at the impact of judge training and feedback on various sources of error in the estimation of the cut-score for a second year medical exam. Three groups of eight judges were provided with 75 exam items and as a training exercise were asked to use an Angoff method to determine the cut-score; subsequently, the judges were provided with examinee item performance data, and asked to modify the item probabilities if they wished. The process of standard setting and use of examinee performance data was repeated with operational items. Using generalizability theory to conduct the analysis, a significant drop in error variance was found for both the training and operational exercises after judges received information on examinee performance and consequential feedback, suggesting that the precision of cut-score estimation can be improved. The results of this study suggest that providing judges with examinee feedback can reduce the variability in cut-scores set by the individual judges; if examinee performance data are not available for the operational items, provision of item performance and feedback in a training exercise for the judges will still have a positive impact on the precision of the estimated cut-score. It was also noted that, on average, across the three groups of judges, the cut-score was raised after judges were provided
with examinee performance data, resulting in fewer examinees meeting the passing grade.

Although numerous studies have focused on the ideal standard setting process, organizations generally do not have unlimited access to judges, items, and funds, yet there is a need to find a balance between available resources and a reasonable standard setting process. In an effort to pursue an efficient standard setting process, Ferdous and Plake (2007) examined the number of items required to estimate a passing score. If a reduced number of items can adequately predict a cut-score, then this would save on the number of hours spent on a standard setting process. Their study suggested that using 40 to 50% of the items to obtain a minimum passing score (MPS) would result in an MPS that was very close to a full-item process. In the case of an exam with 100 items, submitting 40 to 50 randomly selected items to the judging panel did not significantly impact the cut-score, saving time for the judges involved in the process. Ferdous and Plake emphasized that streamlining the standard setting process should not be at the expense of setting accurate cut-scores.

Judge Selection

Judges play an important role in the standard setting process and their abilities have a direct influence on the determination of the cut-score. The panel should consist of enough judges who are adequately credentialed and are trained in the standard setting procedure being used. Criteria for judge selection have been identified by Jaeger (1991), and include attention to domain expertise, and the number of judges. As an example,
domain expertise is an important characteristic if the judges are asked to visualize a minimally competent candidate under the Angoff method. The number of judges chosen impacts the desired (and acceptable) standard error of the mean. Fowell, Fewtrell, and McLaughlin (2008) conducted a small-scale study with medical school exams in an effort to minimize the number of judges with minimal impact on reliability. They found that a reliable pass score could be obtained in one iteration if there were 10 judges, or in two iterations with six judges. This finding identifies that either judges can be decreased or iterations increased without affecting reliability.

It appears, as well, that issues of domain expertise are complex. Plake, Impara, and Potenza (1994) compared passing scores determined by content experts versus non-experts in a study of general educational test scores used to make admission offers to a teacher preparation program. Judges were experts in English, mathematics, science, or social studies, and participated in setting cut-scores in their domain as well as in the other domains. No significant differences were found in the cut-scores derived by content experts and non-experts. Plake et al. (1994) extended this finding to suggest that, in professions with subspecialties such as education or medicine, the judges are not required to be experts in that specific specialty. Verheggen, Muijtjens, van Os, and Schuwirth (2008) also worked to determine the qualifications the judges should have. They used an Angoff procedure but did not provide the judges with the correct answers to the items in order to measure the effect of judges’ knowledge on standard setting. The individual judgment of the items was affected by the judges’ actual subject-related knowledge; it
was found that Angoff estimates were significantly affected by the judges’ abilities to determine the correct answer. Thus, somewhat in contrast to Plake et al., Verheggen and colleagues determined that it is important to have knowledgeable panel members who have some degree of familiarity with the subject matter. Certainly, these findings suggest there is a minimum level of familiarity required by the judges.

In examining a related issue, Verhoeven, Verwijnen, Muijtjens, Scherpber, and van der Vleuten, (2002) examined reliability and credibility when using the same persons as item writers and judges, compared to a panel composed of recently graduated students who did not write the items. They determined that a reliable Angoff estimate required a high number of item writer judges, making this an unfeasible method to use judges. The item writing judges had a higher cut-score than the graduates, and a failure rate of 55% versus a failure rate of 7% with the graduates as judges. As a result, it was suggested that the item writers not sit on the judging panel.

In the case of the flight paramedics, the findings of earlier studies suggest that the panel should be composed of 6 to 10 judges who are not the item writers. If there are physicians on the panel, they need not be transport medicine specialists, but should be knowledgeable about the subject matter.

**Item Development**

Not surprisingly, the quality and type of items used to evaluate the knowledge of the examinees have an effect on the measured scores. Well written items, free of flaws, ensure that the focus of measurement is on construct related knowledge, skills and
abilities, and not on the candidate’s ability to sort through complex and confusing 
questions and answers.

The principles of effective item writing have been well documented (e.g., Case & 
Swanson, 1998; Haladyna, 2004). Nonetheless, improperly prepared items on multiple-
choice examinations have been shown to occur frequently and impact test quality, and 
thereby affect the accuracy of the test results and decisions (Case & Downing, 1989; 
Crehan & Haladyna, 1991; Downing, Baranowski, Grosso, & Norcini, 1995; Frary, 1991; 
examined the effects of flawed items on examinations in medical education, and found 
that multiple-choice items with design flaws reduced the validity of the exam and 
penalized examinees. Examples of flawed questions included those using “all of the 
above,” negative words in the stem, unfocussed stems, and at times “none of the above.” 
In contrast to Downing (2005), Tarrant and Ware (2008) found in a sample of 
undergraduate nursing students (n = 824) that flawed questions were not more or less 
difficult than non-flawed questions. As well, they found that borderline students (those close to the cut-score) benefitted from flawed items since more examinees passed the 
exam when flawed items were retained, as compared to scores calculated when flawed 
questions were eliminated.

The flight paramedic recertification exams in 2006, 2008, and 2010 all consisted of multiple-choice and constructed-response items. While it is generally accepted that 
both types of items are important to use, it is not at all clear how much each should
contribute to the determination of examinees’ scores. Wainer and Thissen (1993) examined how to combine both types of questions into a single score efficiently. They compared two methods of weighting – by component value and by equality. Weighting by equality required both types of questions, multiple-choice and constructed-response, to be of equal reliability. Wainer and Thissen (1993) demonstrated that it is far more costly to create constructed-response items, and more examinee time is needed to achieve reliability equal to the multiple-choice items. A better method of weighting, according to Wainer and Thissen, was that of component value, using criterion weighting. Criterion weighting rescales the points on the test to match the scale of the criterion; weighting would be optimized for maximum predictive validity of the test.

Dependability, Validity, and Decision Consistency

An exam used for certification or licensure is intended to provide the public, the employers, and other stakeholders with a dependable method of demonstrating that examinees have met a standard of performance (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). The exams should be reproducible and consistent (reliable), measure the predetermined competencies (valid), and be able to accurately identify those who fail or pass (decision consistency).

Reliability refers to the consistency of the measurement (Vogt, 2007) and an exam that is highly reliable is considered to be consistent. Reliability is the term used in norm-referenced testing and dependability is used in criterion-referenced testing. An exam is
reliable when repeat administrations of the exam produce the same result, identifying the same students as either competent or not competent. If an exam is dependable, it will correctly differentiate between the categories of performance, and will not result in misclassifications of the examinees. As stated in the Standards for Educational and Psychological Testing, “the need for precision increases as the consequences of decisions and interpretations grow in importance” (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 30). In the case of the flight paramedics, significant decisions are made regarding certification on the basis of the exam results. Hence development of a dependable exam, with accompanying documentation, is required.

Decision consistency estimate methods have been well documented (Berk, 1980; Brennan & Kane, 1977; Hambleton & Slater, 1997; Kolen, Hanson, & Brennan, 1992; Livingston & Lewis, 1995; Traub & Rowley, 1980). Hambleton and Slater (1997) argued that the classification agreement due to chance should be accounted for by using kappa; they also suggest using Subkoviak’s method to estimate decision consistency for a single test. While decision consistency and kappa themselves do not account for error of measurement for candidates near the cut-score, Hambleton and Slater (1997) recommended that error of measurement should be reported for scores near the cut-score. This is supported in The Standards for Educational and Psychological Testing, Standard 2.14 (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999) where “for selection or
classification, the standard errors of measurement should be reported in the vicinity of each cut-score” (p. 35).

Feldt, Steffen, and Gupta (1985) described and compared a number of methods, including item response theory and analysis of variance, to estimate standard error of measurement at specific score levels, and found little difference in estimates from one method to another. They suggested that the researcher choose the method that is most practical to use and most applicable to the study design. As well, we can look to Kolen, Hanson, and Brennan (1992) who described their approach to estimate reliability and conditional errors of scale scores. They suggested that conditional standard errors of measurement can be stabilized, allowing interpretation of examinee scale scores.

The Standards for Educational and Psychological Testing, Standards 3.1 – 3.4 and 3.7, list the requirements to document the test development process, definitions of the test purpose and domain, test specifications such as length, time, and delivery and the procedures to develop, review, and trial the items. The importance of thorough and correct documentation is echoed by Hambleton and Slater (1997): “We would like to see a full reporting by credentialing agencies of decision consistency and kappa for each examination, along with mean examination performance and standard deviation of scores, examination score distribution and the passing score or standard” (p. 36).

Lastly, validity is a major area of concern when developing and evaluating exams or creating a performance standard. According to the Standards for Educational and Psychological Testing, validity “refers to the degree to which evidence and theory
support the interpretations of test scores entailed by the proposed uses of tests”
(American Educational Research Association, American Psychological Association, and
research has been done regarding validity and the consequences of test scores (American
Educational Research Association, American Psychological Association, and National
Council on Measurement in Education, 1999; Kane, 1994; Moss, 1998; Nichols &
Williams, 2009). Validity researchers approach the concept of validity in different ways.
The triarchic model of validity distinguishes three forms of validity evidence: content,
criterion-related, and construct (Crocker & Algina, 1986). Content validity examines how
well the instrument measures what it intended to measure. Criterion-related validity refers
to the correlation of an independent variable with the criterion of interest, and may be
used to predict future performance of the examinee. Construct validity demonstrates an
association between exam scores and the prediction of a theoretical trait (Vogt, 2007). In
contrast, Messick argued that the triarchic model creates an artificial separation of what is
essentially a unified and integrated judgement with consequences of action (Nichols &
Williams, 2009). It is this unified concept of validity that is most commonly used today.
In the case of flight paramedics, the examination of consequential validity could focus on
the impacts of the exam decisions on the paramedics, the MAC, the employer, or patients
as stakeholders. For example, consequential validity would exist if a correlation existed
between the exam score and paramedic field performance that could be assessed through
field assessments or patient complaints.
Generalizability Theory

The analyses in this research were based on the principles of both classical test theory and generalizability theory. Classical Test Theory was developed through the contributions of Charles Spearman, J. P. Guilford, Harald Gulliksen, David Magnusson, Fred Lord and Melvin Novick. The theory is based on the linear equation $\chi = \tau + \epsilon$, where $\chi$ is the observed score, $\tau$ is the true score, and $\epsilon$ represents the error in measurement. The observed score is an estimate of the true score. Assumptions in classical test theory include: the mean of the error scores for a population is zero, the correlation between true and error scores for a population is zero, and the correlation between two error scores on two tests equals zero (Crocker & Algina, 1986). Error and true scores are unique to each examinee, and can only be estimated through the observed score. In the efforts to estimate the true score, the relationship between $\chi$ and $\tau$ can be explored through the reliability coefficient, which is the ratio of variance of the true score to the variance of the observed score.

Limitations of classical test theory, such as the single source of error, led Lee Cronbach and others, such as Robert Brennan, to develop the generalizability theory (G-theory) framework. G-theory is particularly useful for evaluating the dependability of behavioural measurements, especially in performance assessments. In G-theory, the observed score is considered to represent the universe score, and this allows the researcher to generalize from the specific sample to the universe of interest (Brown & Hudson, 2002). More importantly, G-theory allows the exploration of multiple sources of
variance such as person \((p)\), item \((i)\), occasion \((o)\), and rater \((r)\); each of these is defined as a \textit{facet}, or a set of measurement conditions (Crocker & Algina, 1986). \(G\)-theory also allows the determination of the variance components for the interactions amongst the different facets listed above. These variance components can then be used to determine the sources of error that limit the generalizability of the results. For example, in a single facet design, \((p \times i)\), the four sources of variability include differences in examinees achievements, differences in item difficulty, person-item interaction, and unaccounted random error. Variance can be estimated for persons, items, and the interaction between person and item combined with other sources of error.

The sources of error in \(G\)-theory will vary depending on the purpose of the exam. In the case of norm-referenced tests, all variance components that influence the relative standing of the individual contribute to error (Shavelson & Webb, 1991); in a \(p \times i\) design, when persons are the object of measurement, the relative error represents the interaction between persons and items \(\sigma^2_{pi}\). However, the paramedic examinations are criterion-referenced since a single cut score is used to determine certification or recertification based on a defined domain. In criterion-referenced tests, all variance components, except the object of measurement contribute to measurement error.

Credentialing and criterion-referenced testing are areas well suited to the use of \(G\)-theory. Brennan (1992) and Brennan and Kane (1977) applied generalizability theory to credentialing exams, and developed the index of dependability for criterion-referenced testing. Verhoeven, Hamers, Scherpbier, Hoogenboom, and Van der Vleuten (2000) used
$G$-theory to optimize the combination of written and practical components to reduce resources without compromising reliability. Based on a study with medical students ($n = 38$), they showed that reliability increased by adding a written exam to an OSCE. The disadvantage of simply adding another exam is the increased length of time to conduct the composite examination process; however, they argued the practical testing could be shortened by reducing the number of stations, to a point, without diminishing the initial reliability of the assessment process.

Knowledge of score variances can lead to changes in examination practices (Hambleton and Slater, 1997). A decision study ($D$-study) can be conducted within the $G$-theory structure (Crocker & Algina, 1986). A $D$-study uses the information provided in the generalizability study to estimate the optimal designs to meet reliability and dependability requirements. As an example, Hurtz and Hertz (1999) looked at archival data from Angoff workshops with varying numbers of subject matter expert judges to determine the cut-score. Using a $D$-study, they estimated the optimal number of judges to be as few as seven, but ideally 10 to 15. Similarly, Fowell, Fewtrell, and McLaughlin (2008) studied the effects of varying the numbers of items and judges in a standard setting process, using root mean square error to measure the error associated with the cut-score. They found that increasing the number of items does not decrease the error, and determined that six judges are needed if discussion amongst judges occurs.
Existing Standards of Practice for Paramedics

Legislation and regulations govern the paramedic skill set, the acceptable training programs, and the methods of delegation to perform the skills. The National Occupational Competency Profile (NOCP) of Canada explains the provincial jurisdiction related to skill sets and training programs.

The practice of paramedicine in Canada is regulated by each province or, in the case of federal jurisdictions such as the military, by an appropriate federal authority. Each regulator is free to determine the scope of practice and practitioner classification system that applies in its jurisdiction. Similarly the regulator may approve training program(s) that are a prerequisite to employment.

(Paramedic Association of Canada, 2001, p. 3)

The ability for paramedics to legally perform medical acts in Ontario is governed through the College of Physicians and Surgeons of Ontario (CPSO). Because paramedics are not medical professionals designated under the Regulated Health Professions Act (RHPA), they require the delegation of a physician to deliver controlled acts (College of Physician and Surgeons of Ontario, http://www.cpsso.on.ca). The delegating physician, or body of physicians, is obligated to know who is performing the skill, and how the skill level will be monitored. Under Policy 8-10, Guideline 3 states that the delegating physician should “identify the individual performing the act, and be aware of his/her skills.” Guideline 4 addresses competence levels: “establish a process for delegation, that
includes education, ensuring maintenance of competence” (College of Physicians and Surgeons of Ontario, 2010, Guidelines for Delegation section). Legislation and regulations dictate, then, that some form of evaluation must take place, on an ongoing basis; however, specific direction, guidance, or standards from the CPSO are not documented.

The NOCP defines the performance standards that identify specific knowledge, skills, and attitudes for initial paramedic training programs. The National Occupational Competency Profile (NOCP) serves as a reference point for accrediting bodies such as the Canadian Medical Association (CMA). However, the NOCP guidelines can be superseded by an employer or certifying body, and function more as a lower limit of performance expectations. In the case of flight paramedics, much of the scope of practice is above and beyond the definitions found in NOCP, meaning that a performance level has not yet been clearly defined for the purposes of certification exams. Training and certification levels are not standardized across the provinces, and definitely not across nations. For example, in the United States, paramedic certification falls under the National Registry of Emergency Medical Technicians (NREMT), and the National Registry sets the cut-score on the certification exams. Emergency Medical Technicians (EMTs) and paramedics are then licensed by the state in which they are employed.

The frequency, format, and design of the evaluation, in this case the written exam, are decisions left to the Medical Advisory Committee (MAC), a board of physicians overseeing the delegation of controlled acts for the flight paramedics. The MAC has been
using a combination of written assessment, scenarios, and Objective Structured Clinical Examinations (OSCE) as the structure of the recertification process since 2003. A similar model of assessment was examined by Kramer et al. (2002) who found a correlation between a written test (knowledge test of skills) and an OSCE exam ($r = 0.54$) in a study with 47 physician trainees, thus concluding that a knowledge test of skills can be used as an estimate of clinical skills. Kramer et al. (2002) were able to demonstrate that a moderate correlation exists between written and practical methods of testing used in the medical field.

Performance Assessment for Paramedics

The present format for annual certification for approximately 200 flight paramedics was established in 2003 by the MAC, and it provides varying certification levels in Ontario using alternating written and practical exams. Previously, recertification was either an oral exam (1999-2000) or a combined practical and written examination (2001-2003). The examination program has undergone a series of modifications over the past decade in an attempt to deliver an exam that is dependable and results in valid decisions regarding paramedic competence, while balancing the need to avoid excessive manpower or preparation time. There is much value to any educational institution in finding a process that is effective, efficient, and defensible to stakeholders. In turn, there is benefit to the MAC, in terms of increased medico-legal protection against errors, and to patient safety.
The performance requirements for flight paramedics are similar to physicians in scope, skills, knowledge requirements, breadth of knowledge, and decision-making. A key difference is in the legal liability as paramedics work under the umbrella of a physician’s license. Legal responsibilities and accountabilities aside, the nature of the professions is similar enough that comparisons can be made in the research that has been done with physicians and medical students around the world. Recommendations from studies conducted with medical students and physicians can then be applied to the profession of flight paramedics, although future research in the area of assessment and evaluation has the potential for providing valuable information for paramedics of all certification levels.

The Need for Research on Paramedical Examinations

Little research has been done in the area of standard setting, certifications, or criterion-referenced testing in the area of paramedics, pre-hospital workers, emergency medical attendants, or transport personnel. There are many titles used to describe the profession of paramedicine, and the actual title used is dependent on paramedic expectations, the employer, certifying body, or government. A range of skills and knowledge level accompanies the changes in titles as well, so it becomes challenging to compare one workforce of “paramedics” to another. The publications relating to paramedic performance assessment are based on basic pre-hospital skill evaluations. Sheen (2003) looked at developing a practical and comprehensive clinical examination with 36 paramedic students; the Objective Structured Clinical Examination (OSCE)
station exam focused on manual defibrillation skills which is a rudimentary skill in the scope of Ontario paramedics. Perkins et al. (2007) extended the construct to examine cardiac arrest performance in an Advanced Life Support (ALS) scenario with 2449 various health care professionals, and found that there was a significant difference in pass rate, according to the participant’s profession.

The existing research into paramedic certifications is limited, and it is in the interest of public safety to have an examination process that consistently measures performance of flight paramedics. The standard set for the flight paramedics needs to be sufficiently high to assure patient safety and eliminate risk to patient care, yet not so high that there is a shortage of qualified personnel to provide medical air transportation to the citizens of Ontario. Although many studies have been conducted with physician training, additional research is required in the area of certification of flight paramedics to ensure these paramedics have the required skills and knowledge to conduct their work.
Chapter 3

Methodology

Data Collection

My research was based on items and exam results from a written exam administered to 70 Critical Care Flight paramedics in Ontario in the winter of 2010 as part of their annual recertification process. The criterion-referenced high stakes exam consisted of 57 dichotomous multiple-choice and 21 polytomously scored constructed-response items, created through an exam development process. The items were based on specific objectives and a cut-score was determined through an Angoff procedure. Examinations were administered over a three month period in various cities in Ontario, and were proctored by staff in the paramedic education department. Completed exams were sent by courier to a single rater for grading. Item results were entered manually into a spreadsheet. For the purposes of my research, student names were not included in the data. The software used to conduct the analyses included Microsoft Excel, SPSS 17.0, and GENOVA. The Medical Director in charge of the education program granted approval to use these data for the research and provided me with the data set. Ethical clearance was obtained through Queen’s University on December 1, 2010.

During the 2010 exam design process I was part of the team that identified performance objectives, corrected flawed items, and determined the cut-score. After the exam had been conducted, this same team reviewed the performances on the items and the overall exam, and adjusted marks on a few questions that were considered to be
poorly worded. Since the exam was not piloted, questions with unclear wording or
vagueness were not necessarily identified before exam administration. All final decisions
on the exam development process were made by the Medical Director for the education
department and by the Medical Advisory Committee (MAC).

Data Analyses

Descriptive statistics including average exam mark, standard deviation, range, and
number of successful and unsuccessful scores were calculated for the data set. I
conducted item analyses to provide information on those items that supported the correct
classification decisions of the examinees, including dependability and discrimination. An
item is said to be discriminating if there is a high likelihood that high scoring examinees
will get it correct, and a low likelihood that low scoring examinees will get it correct
(Crocker & Algina, 1986). The item analyses included computation of $p$-values, point
biserials, $B$-indices, agreement statistics, and item phi ($\Phi$).

Agreement statistics were computed to look at the probability of agreement
between outcomes on an item and outcomes on the exam (Harris & Subkoviak, 1986),
using the equation:

$$A = 2P_{it} - Q_i - P_t,$$

(1)

where $P_{it} =$ the proportion who passed the exam AND answered the item correctly, $Q_i =$
the proportion who answered the item incorrectly, and $P_t =$ the proportion who passed
exam. $B$-index values were calculated for the dichotomous items in order to determine
how well the item contributed to the pass/fail decision:
where \( \text{IF}_{\text{pass}} \) = the item difficulty for those examinees who passed the exam and \( \text{IF}_{\text{fail}} \) = the item difficulty for those who failed the exam.

\[
B\text{-index} = \text{IF}_{\text{pass}} \cdot \text{IF}_{\text{fail}} 
\]

Phi (\( \phi \)) was used with the dichotomously scored items to correlate item performance to overall exam performance, and was calculated as:

\[
\phi = \frac{(P_i - P_t) \cdot \sqrt{P_i \cdot Q_t}}{P_i} 
\]

where \( P_i \) = the proportion who answered the item correctly and \( Q_t \) = the proportion who failed exam. Although the \( B \)-index and \( \phi \) give essentially the same information, the \( B \)-index will be greater than \( \phi \) when the exam variance is greater than the item variance (Hudson, 1993).

In an attempt to examine whether a relationship existed between actual \( p \)-values and those predicted by the standard setting panel, the differences between predicted \( p \)-values (\( p_{\text{pred}} \)) and actual \( p \)-values (\( p_{\text{actual}} \)) were calculated by subtracting the calculated \( p \)-value from that predicted by the standard setting panel. The correlation was then calculated between \( p_{\text{pred}} \) and \( p_{\text{actual}} \) for all questions together, multiple-choice items alone, and constructed-response items alone.

Because criterion-referenced testing is not concerned with variability amongst examinees, most norm-referenced reliability equations are not applicable. Consistency, or dependability, is measured in criterion-referenced tests by examining classification consistency. Most consistency calculations assume more than one administration of an exam, as a test-retest or parallel form. Huynh developed a gold standard approach to
estimation with a single test administration, but this method has been found to be more complicated than the simpler approximation developed by Peng and Subkoviak in 1980. This approximation incorporates the agreement coefficient ($\rho_0$) and the kappa coefficient ($\kappa$). This estimate is computed by first calculating the standardized $z$-score for the cut-score as:

$$z = \frac{c - \mu - 0.5}{\sigma} \quad (4)$$

where $c =$ cut-score, $\mu =$ average score of the exam for all candidates, and $\sigma =$ standard deviation of exam scores. Although equations exist to calculate $\rho_0$ and $\kappa$, tables developed by Subkoviak (1988) can be used as an approximation in place of more difficult calculations. The agreement coefficient, ($\rho_0$), can be determined using the $z$-score from equation 4, and the reliability of the exam scores, as approximated by the Kuder-Richardson coefficients, KR-20 or KR-21 (Subkoviak, 1988). Similarly, the kappa coefficient ($\kappa$), can be approximated from Table 3 (Subkoviak, 1988), using the same $z$-score value and KR-20 or KR-21. In this study, $\rho_0$ and $\kappa$ were calculated for the dichotomous data only. Although both KR-20 and KR-21 were estimated, KR-21 is used in subsequent calculations, as this formula assumes that the items have varying difficulties.

**Generalizability Theory**

Under generalizability theory, $G$-theory, all calculations were done with the dichotomous items. The cut-score for the dichotomous data was recalculated, using the raw data documented during the standard setting process, and was determined to be 45/57
(= 78.9%). The Brennan-Kane index, Φ(λ), and the dependability coefficient, (Φ), reflecting the contribution of the measurement procedure to the criterion-referenced dependability were calculated. Signal-noise ratios (S/N) were calculated by the ratio of universe score to error variance, where S/N for Φ = σ²(ρ)/σ²(Δ) and S/N for Φ(λ) = [σ²(ρ) + (μ - λ)²]/σ²(Δ), providing an index of the relative precision of the measurement procedure (Brennan, 1992). A 95% confidence interval for the universe score, μₚ, was estimated in order to describe the bounds of the universe scores, under the assumption that absolute error scores were normally distributed. G-theory provides a mechanism to examine the sources of error in a criterion-referenced exam; sources of error in this p x i study are the differences across the mean scores, σ²(ᵢ), as well as the person-item interactions, σ²(ᵢₚ). Variance estimates, under G-theory, were done for σ²(ᵢ), σ²(ᵢₚ), and σ²(ᵢₚᵢ) to determine σ²(Xᵢₚ), the variance of Xᵢₚ, where Xᵢₚ is the score for a person p on item i. Through the use of a decision study (D-study), these variance components were used to explore recommendations for various exam designs. Given the single facet design, the D-study was limited to estimating the number of items that would be required to obtain specified levels of decision consistency in future examinations.

Interview with Medical Director

Along with the quantitative data analysis, the Medical Director for the education department was interviewed to collect feedback on the revised process and the results of the 2010 exam, and to provide suggestions for improvement in future years. This feedback was used to address the research question regarding recommendations for future
exams. It was important that the recommendations be practical and feasible for the MAC, and that they be capable of implementing the recommendations should they choose to do so. Interview questions (Appendix D) covered topics such as desired pass rates, item difficulty, and exam length. Questions were asked about the development process used in 2010 and we discussed the exam development for the upcoming year.

The interview was conducted by telephone at a time that was agreed upon by both parties in January 2011. The Medical Director was not provided with the questions beforehand, although he was aware that I was going to ask his opinions and views on the 2010 exam, and future exams. We previously had held many conversations about the 2010 exam development and administration, and the Medical Director was closely involved in all aspects of the 2010 exam; however, I did not use any of the content of these conversations and meetings in this research. Notes were taken and recorded during the interview. The questions were posed in an open-ended format so as to promote deeper discussion and allow for exploration of thoughts and ideas. The main intent of this interview was to gather information for operational purposes. The Medical Director’s answers provided a foundation to create recommendations for future exams that were realistic, practical, and achievable. Standards which are unrealistically high may not be achievable, and I felt that the final recommendations should be reflective of the published research and documentation, and interview feedback.
Chapter 4

Results

This chapter presents the findings of various analyses performed with the data from the 2010 exam. Descriptive statistics are presented for the 2010 exam and compared with those from 2008. A section on item analysis of dichotomously and polytomously scored items, along with a description of the problematic items follows. The next section covers the results of the generalizability analyses (G- and D-studies) using generalizability theory as a foundation for analysis. Decision consistency and dependability indices are provided, and finally qualitative information gathered during the interview with the medical director is summarized.

Descriptive Statistics

I obtained the data set for the 2010 recertification exam as a Microsoft Excel spreadsheet which included exam and item scores for Critical Care paramedics; all candidate names were removed to ensure confidentiality. The data were screened for inconsistencies; in the case of an invalid score, the paper copy of the paramedic’s exam was reviewed to acquire the correct item value.

A comparison of descriptive statistics for both the 2008 and 2010 exam, as provided in Table 1, illustrate that the percentage of examinees passing the exam was the same in both years, but the average score in 2010 ($\bar{x} = 81.1\%$) was higher than in 2008 ($\bar{x} = 73.3\%$). The cut-score in 2008 was set at 70%, and in 2010 at 77%.
Table 1

Summary of Descriptive Statistics for Recertification Exams for 2008 and 2010

<table>
<thead>
<tr>
<th>Year of Exam</th>
<th>2008</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of examinees</td>
<td>82</td>
<td>70</td>
</tr>
<tr>
<td>Total number of items</td>
<td>90</td>
<td>78</td>
</tr>
<tr>
<td>Number of dichotomously scored items</td>
<td>79</td>
<td>57</td>
</tr>
<tr>
<td>Average total score</td>
<td>73.3%</td>
<td>81.1%</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>5.6</td>
<td>4.9</td>
</tr>
<tr>
<td>Range</td>
<td>68.5-98</td>
<td>69.9-89.1</td>
</tr>
<tr>
<td>Percentage of successful candidates</td>
<td>74.4% (n=61)</td>
<td>74.3% (n=52)</td>
</tr>
<tr>
<td>Percentage of unsuccessful candidates</td>
<td>25.6% (n=21)</td>
<td>25.7% (n=18)</td>
</tr>
<tr>
<td>Cut-score (%)</td>
<td>70</td>
<td>77</td>
</tr>
<tr>
<td>Standard setting process</td>
<td>None</td>
<td>Angoff</td>
</tr>
<tr>
<td>Item review process</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Figure 1 provides the distribution of scores across the 70 examinees. The scores were negatively skewed, with a majority of scores to the right of the cut-score of 77%. There were a small number of scores far to the left of the cut-score, reflective of the larger proportion of examinees passing the exam. The frequency distribution curve for criterion-referenced exams is not expected to be that of a normal distribution; rather,
because it measures mastery of a predetermined performance standard, it may be clustered and skewed.

![Frequency Distribution for Exam Scores](image)

**Figure 1** Frequency Distribution for Exam Scores

**Item Analysis**

The exam consisted of 57 dichotomously scored multiple-choice and 21 polytomously scored constructed-response questions. Because some analyses can be done only with dichotomously scored items, certain analyses involved all items, and others isolated the dichotomously scored items.

The correlation between predicted $p$-values and actual $p$-values for all items as seen in Figure 2 illustrates that the panel judges’ predictions were positively correlated to the actual $p$-values with a Pearson correlation of $r = 0.97$. 

40
Figure 2 Correlation between Predicted and Actual $p$-values for All Items

Separating the multiple-choice items from the constructed-response items, Figure 3 illustrates the Pearson correlation between the actual and predicted $p$-values for the multiple-choice items, found to be $r=0.55$. The correlation between the predicted $p$-values and actual $p$-values for the constructed-response items was considerably higher than the multiple-choice items with $r=0.94$, as seen in Figure 4.
Figure 3 Correlation between Predicted and Actual $p$-values for Multiple Choice Items

Figure 4 Correlation between Predicted and Actual $p$-values for Constructed Response Items
The relationship between the items and the examinee performance was examined through $B$-indices and agreement statistics, and calculated for all dichotomously scored items. These values can be found in Appendix E.

Point biserials ($r_{pb}$) were calculated as a measurement of discrimination. The $r_{pb}$ values for the dichotomous items ranged from -0.12 to 0.48. No point biserial correlation was calculated for five of the items; three items were answered correctly by all and examinees were granted full marks for two items that were determined, post-exam, to be poorly worded. Three items had negative point biserial correlations, indicating these questions did not discriminate masters from non-masters. There were no questions with zero correct responses.

Problematic and Discriminating Items

One purpose of item analysis is to identify those items that should be retained for future exams, and to identify those that appear to be problematic. A problematic item may be disregarded in the examinee’s final score, or may become useful in future exams with some revisions. There are no black and white rules to identify problematic items, but an item should discriminate well, as indicated by point biserial values that are ideally greater than 0.25, and there should be some agreement between those who answer the item correctly and those who pass the exam. Criterion-referenced items may not always be as highly discriminating as norm-referenced ones, so it may be acceptable to use items with point-biserial values of less than 0.15. The knowledge assessed by the item should
be considered, along with the item analyses, before a decision is made to retain an item for future use, discard it, or modify it.

Items were analyzed and then differentiated as discriminating, problematic, or neither. Items were categorized into those that discriminated well ($r_{pb} > 0.25$, $n = 21$), were borderline ($0.15 < r_{pb} < 0.25$, $n = 9$), or discriminated poorly ($r_{pb} < 0.15$, $n = 22$). The analyses for the 21 items that discriminated highly are presented in Appendix F. Of these highly discriminating items, there were 10 instances where the predicted $p$-value was 20% greater than the actual $p$-value. Although no literature exists to define criteria for problematic items, I flagged items as being problematic for either or both of the following reasons: 1) a point biserial correlation, $r_{pb}$, < 0.15 including any negative values, 2) an actual $p$-value that differed by more than 20% from the predicted value. Over prediction indicates that many examinees performed below the expected performance level of the minimally competent candidate, and may flag areas for future education; however, over prediction may also be an indicator that the judges involved in setting the cut-score may not be using an appropriate definition of a minimally competent candidate. All items were related to a performance objective, making it possible to identify the categories where the $p$-values were greater than the actual $p$-values. There were three areas that were predominantly over-estimated: cardiovascular, neurology, and pediatrics, implying that the panel believed the paramedics knew more about these constructs than they demonstrated on the exam questions. Hence it may be important to consider these topics for future education. The item analysis values for the problematic
questions are found in Appendix G; three items in this group had negative point biserial correlations.

**Generalizability Theory**

All 57 dichotomously scored items were analyzed using generalizability theory (G-theory) in a single-facet crossed study of persons (p) crossed with items (i). The estimated error variance components for persons, items, and persons crossed with items are presented in Table 2. The variance component for persons accounted for just over 1% of variance ($\sigma^2(p) = .0029$), corresponding to the largely similar scores obtained by the examinees. The variance component for items was almost 31% ($\sigma^2(i) = 0.0636$), and for persons crossed with items was 67.7% ($\sigma^2(pi) = 0.14$). The high estimated variance component for persons crossed with items is an indicator that almost 2/3 of the variability (random error) lies within this relationship, and provides an estimate in the changes in the relative standing of a person from item to item.

Total variance $= \sigma^2(p) + \sigma^2(i) + \sigma^2(pi) = 0.21$. The absolute error variance, $\sigma^2(\Delta)$, which reflects the error involved by using an observed score to estimate a universe score, was 0.0036, and the estimated absolute standard error of measurement, $SEM$, was 0.06. The dependability coefficient, $\Phi$, an index that reflects the contribution of the measurement procedure to the dependability of the exam, was calculated to be 0.446. This suggests that the variance in decisions were nearly 45% dependable, and 55% undependable. The Brennan-Kane index of dependability, $\Phi(\lambda)$, using the cut-score for
only the dichotomous items of 79%, was 0.99; Φ(λ) relates to the degree to which classifications are consistent.

Signal-noise ratios, S/N, were computed for Φ and Φ(λ) for the dichotomous items, and were found to be 0.89 and 3.04, respectively. If the signal is strong compared to the noise, the appropriate discriminations can be made. The signal should be stronger than the noise, so a desirable ratio would be anything greater than one, i.e. S/N > 1. The signal-noise ratio estimated for Φ was less than 1, so some intended discriminations of the exam have been lost (Brennan, 2003). The signal-noise ratio for Φ(λ), where λ = 79% for the dichotomous items, was more than 3 times as large as the Brennan-Kane index of dependability, indicating strong discrimination at the cut-score. Good measurement practice, in general, should include maximization of noise reduction (van der Vleuten, 2010). The 68% confidence interval for μ_p, the universe score, was estimated to be between 64 and 76, indicating that 68% of the true scores lie within this range.

Table 2

*Estimated G- Study Variance Components from 2010 Recertification Exam*

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>Estimated Variance Components</th>
<th>Percentage of Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>69</td>
<td>20.89</td>
<td>0.31</td>
<td>0.0029</td>
<td>1.4</td>
</tr>
<tr>
<td>i</td>
<td>56</td>
<td>257.04</td>
<td>4.59</td>
<td>0.064</td>
<td>31.0</td>
</tr>
<tr>
<td>p,i</td>
<td>3864</td>
<td>537.02</td>
<td>0.139</td>
<td>0.14</td>
<td>67.7</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.21</td>
</tr>
</tbody>
</table>
Decision Study

A decision study allows us to generalize from observed scores to a universe score. I was interested in looking at the numbers of dichotomous items, specifically multiple choice questions, that would generate different levels of dependability, balanced with the resources required to develop and administer future exams. Table 3 illustrates the changes in absolute error variance and dependability as we change the number of dichotomous items from 25 to 200, in varying increments. At 200 dichotomous items, the index of dependability (Φ), considering only the dichotomous items, approaches 0.7411, indicating 74% dependability.

Table 3

*Decision Study Results*

<table>
<thead>
<tr>
<th></th>
<th>G study</th>
<th></th>
<th>D study</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n= 57</td>
<td>n=25</td>
<td>n=50</td>
<td>n=75</td>
<td>n=150</td>
<td>n=200</td>
<td></td>
</tr>
<tr>
<td>(\sigma^2(p))</td>
<td>0.0029*</td>
<td>0.0029</td>
<td>0.0029</td>
<td>0.0029</td>
<td>0.0029</td>
<td>0.0029</td>
<td></td>
</tr>
<tr>
<td>(\sigma^2(i))</td>
<td>0.0636</td>
<td>0.0025</td>
<td>0.0013</td>
<td>0.0009</td>
<td>0.0004</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>(\sigma^2(pi))</td>
<td>0.139</td>
<td>0.0056</td>
<td>0.0028</td>
<td>0.0019</td>
<td>0.0009</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>(\sigma^2(\Delta))</td>
<td>0.0036</td>
<td>0.0081</td>
<td>0.0041</td>
<td>0.0027</td>
<td>0.0014</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td>(\Phi)</td>
<td>0.446</td>
<td>0.2635</td>
<td>0.4171</td>
<td>0.5177</td>
<td>0.6822</td>
<td>0.7411</td>
<td></td>
</tr>
</tbody>
</table>

*Note. Due to the small absolute values, the values in this table are listed to 4 decimal places.*
Decision Consistency and Reliability

Cronbach’s $\alpha (\alpha = 0.65)$ is an internal estimate of consistency and was calculated using the dichotomous and polytomous data. The value from this data set indicated that at least 65% of the observed variance for the whole exam can be attributed to true variance, and implied that several attributes may have been measured by this exam. Using the Peng Subkoviak method for determining decision consistency in a single exam administration, the index of consistency for the 57 dichotomous items, $p_o$, was estimated to be 0.85 using KR-20. Thus 85% of the time, the classifications were considered consistent and 15% of the time they were inconsistent. The desired value for $p_o$, is, according to Subkoviak, dependent on the seriousness of the decisions being made from the exam results (Subkoviak, 1988). Subkoviak suggests that high-stakes exams should be sufficiently long to result in $p_o > 85\%$.

Kappa compares classification consistency to random assignment by chance. The closer the cut-score is to the middle of the distribution, the higher the value of $k$. The calculated value of kappa ($\kappa = 0.26$), using KR-20 means this exam has achieved an increase over chance classification by over 26%. The ideal kappa, as with the index of consistency, depends on the intent of the exam results. Typically, an exam with 10% of examinees not demonstrating competence should have a $\kappa = 0.45$. In the case of this study, the kappa indicates the exam is somewhat dependable, but not as high as it should be for a high-stakes exam (Subkoviak, 1988).
Interview Feedback

The Medical Director of the education program, who acts as the link between the education department and the Medical Advisory Committee (MAC), was interviewed to collect qualitative feedback on the exam development process and the results. The interview questions were a guide for the discussion for the purpose of this research, but opportunities existed to expand on or clarify intentions of questions if necessary. Because we had previously had discussions, through the nature of our work together, it was important to reiterate earlier decisions and opinions through systematic questioning. The Medical Director’s main areas of concern with the 2010 exam included: weighting of content, the length of time to analyze and process results, inadequate training for the item writers, and minimal time to prepare items and exam for administration. With respect to development of an exam item bank, he stated: “The bank needs to be sufficiently large to avoid memorization of all questions; in the ideal world there would be about 500 questions. MAC views this as important.” (M.D., personal communication, January 30, 2011).

He felt that the overall approach in 2010 was grounded in methodology and that MAC felt more secure in delegating to the paramedics after they had been recertified with this exam. We discussed the potential impact of reverting back to the “old” system of exam development and he stated: “This would be a disaster. It would undermine the credibility of our new system. Stepping back would decrease our confidence that the public is being protected.” (M.D., personal communication, January 30, 2011)
According to the Medical Director, inadequate consistency in classification was not a cause of worry, as those paramedics who were unsuccessful underwent a remediation process that included a discussion of constructs with a Medical Director from MAC. During these discussions, areas of weakness could be teased out and corrected, and items which did not adequately test the paramedic’s knowledge level could be identified for future modifications. The 2010 exam generated more discussion between paramedic and physicians, and MAC felt that this exam was more reliable and had improved content validity, based on their impressions (M.D., personal communication, January 30, 2011).

In the interest of practicality, the recommendations needed to be achievable in an operational environment. Although the information collected was based on an interview with a single person, this Medical Director was highly involved in all phases of the 2010 exam and was the most knowledgeable member of MAC with respect to this topic. This interview played an important role in creating the foundation for the recommendations made for future exams. The MAC was interested in administering an examination that was more defensible and credible. Yet, impractical or unachievable recommendations would not assist MAC with improving the quality of the exams.
Chapter 5

Discussion

This study was based on three research questions focused on a written recertification exam for flight paramedics, and the impact of changes made to the development process in 2010. An unusually high failure rate in 2008 led the Medical Advisory Committee (MAC), in their role as the certifying body, to look at and make improvements to the methodology of content development and standard-setting. The discussion is presented in five sections: discussion of research findings related to each of the three questions, limitations, and the conclusion.

Research Questions

Question 1: What systemic approaches were used in the 2010 paramedic recertification exam process and how did these changes potentially impact the measured scores?

In this section, the changes made to the 2010 exam are explained and the potential impacts of these changes are discussed. The process of creating and administering a high-stakes exam is complex, and relies on many people and a variety of roles to run smoothly. Hambleton and Slater (1997) described the “complex interactions among exam scoring, standard setting, and the reliability and validity of pass-fail decisions” (p. 20). Many improvements were made to the 2010 exam, throughout the development, administration, and analysis phases. Objectives for the examination were set according based on the National Occupational Competency Profile (NOCP) for paramedics in Canada. Items were reviewed to ensure they were clearly written, and flaws in the stem
and leaf were corrected. A standard setting process was followed to determine the passing score. The examination was proctored and a post-exam item analysis was conducted.

**Objectives**

High-stakes exams, such as the paramedic recertification exams, should be criterion-referenced, as this is the most appropriate test of competence (Norcini, 2003). Criterion-referenced exams measure the examinees' knowledge as compared to an identified set of behaviours or objectives required by the profession; candidates who meet a minimum standard of performance on the exam are considered to be successful. The 2010 paramedic recertification exam was based upon objectives from the NOCP guidelines defining the professional competencies required by all paramedics in Canada, with a performance standard set through an Angoff process. The objectives were drawn from this profile with approval by MAC, and examination items were created to specifically address these objectives. Previously, exams were based on broadly defined body systems such as cardiology, and little attention was actually given to the requirements of how the examination items measured the knowledge and skills paramedics would be expected to demonstrate in the field. Perusal of previous exams revealed items that were vaguely worded and not easily linked to specific objectives. Thus examinees’ scores may not have been indicative of each examinee’s ability to meet the requirements of the profession. The Medical Director felt that the examinee scores on the objectives-based exam, such as the 2010 exam, provided more accurate information.
on the knowledge levels of the paramedics (M.D., personal communication, January 30, 2011).

**Item quality**

Item quality is important in high-stakes testing since poorly constructed items can differentially affect the scores of individual and subgroups of examinees, leading to misclassifications (Downing, 2005; Tarrant & Ware, 2008). Flawed items, according to Downing (2005) can lead to an increased failure rate, and subsequent false classifications. In 2010, the paramedic examination items were written by writers with expertise in the subject matter, based on specified objectives. Items were modified to eliminate potential flaws, such as an “all of the above” option in the distractors, or a negation in the stem. The efforts put into removing item format flaws on the 2010 examination likely resulted in a more accurate classification of examinees who wrote this exam.

**Standard Setting process**

In previous paramedic recertification exams, the cut-score was set arbitrarily at 70%. While consistent and familiar to the examinees, the 70% cut-score did not reflect the changing difficulties of the examinations across administrations, nor differentiate between those who demonstrated competency and those who did not. The 2010 paramedic recertification exam represented the first time that a standard setting procedure was used to create a cut-score. The Angoff method was selected, recognizing that other methods do exist and that these methods all have their strengths and weaknesses (Cohen-Schotanus & van der Vleuten, 2010; Kane, Crooks, & Cohen, 1999; Muijtjens,
Hoogeboom, Verwijnen, & van der Vleuten, 1998; Nasstrom & Nystrom, 2008). In 2010, the standard setting panel included two physicians who were members of MAC, and two teachers who taught in the paramedic education department and were certified as critical care paramedics. The acceptable performance level was now identified through a recognized process that determined the cut-score to be 77% for all items, or 79% for dichotomous items only, increasing the confidence in the resulting examination decisions. Although the percentage of examinees failing the exam in 2010 did not change in comparison to 2008, the classification decisions were now based on a defined standard. Nonetheless, the first effort at creating standards for this exam was challenging. For example, only four judges completed one iteration of the standard setting process, suggesting the process was under-resourced. I was the facilitator of the standard setting process, and had no previous practical experience in leading the group through the process. Hence it is also possible that the accuracy and reliability of the selected cut-score may have been compromised by an inexperienced facilitator and low numbers of judges.

Administration of the Examination

In 2008 examinees wrote the paramedic recertification examination on-line, in an unsupervised setting, raising concerns about the accuracy of these results. The unsupervised on-line exam led to opportunities for examinees to consult resources such as reference textbooks or co-workers. In 2010, the examination was written with pen-and-paper because the on-line format was no longer available. The examinations were proctored, reducing the possibility that examinees used non-permissible resources.
Secondly, the 2008 examinations were scored on-line whereas the 2010 examinations were sent to a single rater for grading. Once graded, the data were entered into a spreadsheet and this transcription from paper into a spreadsheet increased the potential for coding errors. When the data were checked and cleaned, a number of coding errors were found, such as item scores recorded with values higher than the maximum score. The transition to paper also caused a delay between the time that the examination was written and graded, due to shipping time and the time for grading. The Medical Director was concerned about the prolonged time to summarize and report results and the time it took to conduct item and exam analyses. Since the paramedics work under a delegatory model, the Medical Advisory Committee felt that patients were potentially at risk if failures were not reported promptly. Using an automated scoring system, where examinees write the exam online, would allow for automated scoring and more timely analyses.

Item Analysis

After all examinees had completed the exam consisting of 57 multiple-choice and 21 polytomous constructed-response questions, results were collated and the items were reviewed by the Medical Director and myself. Upon review, we felt that three items were poorly worded and all candidates were granted full marks on these items. Item scores for each item were entered into the spreadsheet and the item discrimination indices and difficulties were calculated. This information will be valuable in determining which questions to use again in future exams; highly discriminating items, e.g. point biserial
value > 0.15, can be used in the 2011 exam, and retained in an item bank. Only 21 of the
dichotomously scored items in the 2010 examination were highly discriminating with
point biserial values of > 0.25 and 22 were poorly discriminating. While negatively
discriminating items should not be used in the future, criterion-referenced exams often
result in items that are not as discriminating as those in a norm-referenced exam. Given
that the purpose of the examination is to correctly classify paramedics’ overall
competency, the number of poorly discriminating items makes it difficult to ensure that
consistent decisions are being made. However, the signal-noise ratio for $\Phi (\lambda)$ leads us to
believe that the intended discriminations can be made around the dichotomous cut-score
of 79%. In future examinations, efforts should be focused on increasing the
discrimination of examination items.

Once again, there are opportunities to further these analyses. It was not possible to
determine some of the item statistics in detail for the 2010 examination. For example, the
proportions of examinees choosing the different incorrect distractors were not recorded.
Such analyses could provide further opportunities to improve item development,
including the wording, content, or meaning.

Impact of Changes

Despite the enhancements in the areas discussed above, there was no change in
the failure rate from 2008. Changes to the process did not appear to lead to changes in
the proportions of paramedics passing the 2010 administration of the certification
examination. At one level, the lack of apparent improvement may be worrisome;
however, one must be cautious when comparing the 2008 results to those of 2010. While the overall results did not change, the development, administration, and standard setting activities associated with the 2010 proctored exam likely provide a truer reflection of the paramedics’ knowledge and understanding with respect to the expectations on the examination. Many of the areas of change made in 2010 have been shown in the literature to improve dependability and decision consistency.

**Question 2: What was the decision consistency for this exam, and what were the factors that impacted this consistency?**

Decision consistency is the degree to which different parallel forms of an examination will result in the same examinee scores (Crocker & Algina, 1986). It is challenging to administer parallel high-stakes exams, so decision consistency must often be calculated from a single administration. One method of determining decision consistency in a single administration was developed by Peng and Subkoviak. Using this method to calculate $p_0$ and kappa ($\kappa$) for the dichotomous items only, the results for the 2010 administration of the paramedic recertification examination were found to be $p_0 = 0.85$ and $\kappa = 0.26$. This value of $p_0$ indicates that decisions could be considered consistent 85% of the time. One expects in a high-stakes exam where human lives – those of patients – are at stake, the classification of those expected to treat such lives is greater than 85% (Subkoviak, 1988), and the results from the 2010 can be considered, according to these calculations, to fall within acceptable limits. Similarly, the $\kappa$ value indicates that 26% of classifications are consistent beyond the percentage expected by chance for this
recertification examination. One expects kappa to equal 0.45 ($\kappa = 0.45$) (Subkoviak, 1988), and the 2010 data do not meet this value. Kappa is affected by length of the exam and variability of examinees, so increasing the number of items or having greater variability amongst examinee scores will lead to an increased kappa value.

The dependability of this exam, a further measure of classification consistency, can be measured according to generalizability theory, by determining the generalizability or dependability coefficient ($\Phi$). According to Brennan (2003), $\Phi$ should be > 80%; however, in this exam, it was calculated to be 0.446, meaning that this exam was 45% dependable. An important advantage of $\Phi$ is that it can be used to determine the sources of error that reduce classification accuracy and the methods to best improve such classifications. Based on the formula, $\Phi$ can be increased by reducing absolute error.

Absolute error is dependent on the variance from items, $\sigma^2(i)$, variance from persons interacting with items, $\sigma^2(pi)$, and the number of items. The majority of error variance for this examination was due to interaction of persons with items ($\sigma^2(pi) = 67.7\%$); lowering this variance would lead to an increase in the dependability. As well, the number of items on the exam could be increased from the 57 dichotomous items to result in a decreased absolute error, and thus, increased dependability index. The decision study was used to determine the ideal number of items that should be on subsequent examinations. Overall, a 75-item test with 57 dichotomous items, resulted in a low dependability of 45%. Increasing the number of items results in gradually increasing dependability, with 200 dichotomous items resulting in an estimated dependability of 74% ($\Phi = 0.7411$).
However, creating items for such a lengthy exam will be costly and time consuming, and decisions will need to be made within the MAC as to an acceptable dependability level. The dependability index, $\Phi (\lambda)$, where $\lambda =$ cut-score of 79% for the dichotomous items, was 0.999. This index describes the agreement in decisions based on the cut-score. One should keep in mind that $p_o$, kappa ($\kappa$), $\Phi$ and $\Phi (\lambda)$ do not reflect contribution to the examinees scores by the constructed response items.

The estimates of decision consistency and dependability were varied, due to some degree of over or underestimation inherent in the calculations, and due to the methodology of calculations. For example, $p_o$ ($p_o = 0.85$), based on threshold-loss, tends to overestimate decision consistency because it is based on only one exam administration and looks at the degrees of mastery. On the other hand, $G$-theory looks at sources of error variance; the coefficient, $\Phi (\Phi = 0.446)$, based on squared-error loss, is a general purpose estimate of dependability of the domain score. The estimates for calculations for decision consistency and dependability in this study are based on dichotomous items only; because the polytomous items were not incorporated into the analysis, it is likely that these values are underestimated.

Certainly, improving item quality could also increase dependability and developing a solid process for item creation should decrease the item error variance. Nevertheless, it is important to continue to examine the dependability of subsequent examinations that are based on a single index or coefficient in $G$-theory, as such examinations may not be reflective of the body of such examinations used in this context.
Similarly, it is important to not focus all of the improvement efforts on one single source of error variance. Rather, all error variances and variance components should be taken into account to make judgements and decisions in $G$-theory environments (Brennan, 2003). As an example, van der Vleuten (2010) argues that the largest source of noise (error) is usually exam difficulty.

Associated with dependability, poor item quality has an impact on the discrimination of items. In the case of the paramedic recertification examination, the signal-noise ratio for $\Phi$, ($\Phi = 0.89$) suggests that the examination items come close to discriminating between examinees with different skill levels, although a value $> 1$ should be a target value. The point biserial correlations across the items provide further evidence of challenges with test discrimination.

*Question 3: What recommendations can be made to develop a more consistent dependable exam in the future?*

The final research question was intended to identify recommendations to increase consistency and dependability for future paramedic recertification examinations. Combining the quantitative results from the examination results and qualitative feedback from the interview, the three categories of recommendations are standard setting, item and exam development, and exam administration. Recommendations were based on data collected in this study, findings from other published studies, and best examination practices.
Standard setting Process

One of the critical changes made to the 2010 administration of the paramedics’ recertification examination was the introduction of a standard setting process. However, given that this was the first time the process was ever completed for this assessment, several issues and challenges were identified. Based on these challenges, four recommendations are provided with respect to subsequent standard setting procedures.

1. Increase the number of judges participating in the standard setting process as supported in the literature from the four used in 2010, to up to 10 judges, to improve the reliability of the cut-score. (Fowell, Fewtrell, & McLaughlin, 2008; Hurtz & Hertz, 1999; Jaeger, 1991). In situations where a lack of manpower and costs are of concern, the number of judges can be reduced and more iterations performed. If time constraints are an issue, it is possible to reduce the number of items the judges review by half (Ferdous & Plake, 2007). Hence MAC has the option of a) using fewer than 10 judges but adding iterations when manpower is limited, or b) using 10 judges but reviewing only half of the items in the standard setting process, when time is limited.

2. Conduct training for the judges. To reduce the error in determining the passing mark, literature supports the need for adequate judge selection, training, and monitoring (Bandaranayake, 2008; Clauser, Swanson, & Harik, 2002). For example, Plake (2006) recommends that the panel practise on items similar to the proposed exam items. In the 2010 process, only limited written instructions were
given to the judges, and no time was allocated to the panel to discuss or clarify the process or answer questions.

3. Ensure the standard setting panel consists of carefully selected judges. In the interview, the Medical Director stated that he believed that since paramedics work under the license of a physician, a physician should determine what the paramedics need to demonstrate competency (M.D., personal communication, January 30, 2011). However, physicians do not perform the work of paramedics. Hence the judging panel should include certified flight paramedics who understand the expectations of the profession. As seen in other studies, physicians who are judges should have various levels of expertise and specialty but should not be the item writers (Jaeger, 1991; Fowell, Fewtrell, & McLaughlin, 2008; Verhoeven, Verwijnen, Muijtjens, Scherbier, & van der Vleuten, 2002).

4. Continue to use the Angoff procedure, often considered the gold standard for standard setting. At the same time, examine other increasingly used standard setting procedures. Potential options include the borderline group technique, the borderline regression method, the Hofstee method, the Nedelsky method, or the contrasting groups method.

*Item and Exam Development*

Another area of change in the 2010 administration concentrated on improving item qualities such as discrimination and wording. Item analysis relied on a single person to perform the psychometrics, with minimal software access. The next four
recommendations are made to ensure sustainability of this area, and that future improvements to items continue to be made.

1. Items need to be developed that are discriminating and have few if any flaws.

   Items should ideally be piloted before they are used in an exam. The items to be piloted can be included in the exam, but are not included in the examinee’s score. These pilot items are then used for analysis only, and may be retained for future use, rejected, or revised and re-tested. Data in this study show that the correlation between $p$-values predicted by the judges and actual values was lower on the multiple-choice items ($r = 0.55$) than on the constructed response items ($r=0.94$). Multiple-choice items are the most difficult to develop, and the constructed-response items appear to be offer valuable contributions to the exam. Hence, a combination of both formats should continue to be included in future exam designs.

2. A larger bank of items needs to be developed for future use. In the 2010 exam the item variance accounted for 31% of the variability in scores, and only 21 of the 57 dichotomously scored items were strongly discriminating.

3. Invest in a software package to perform the item analysis. As the examination process moves forward, a consistent methodology for item and test analysis, as seen in best practice, will need to be established. Commercial packages are available for purchase at affordable cost. Some options include jMetrik (itemanalysis.com), an open-source software that can produce frequency analysis,
descriptive statistics, point biserials, and calculate decision consistency. Iteman™, offered through Assessment Systems Corporation, includes the same features as jMetrik, but also offers test development and delivery software. Lertap 5 provides psychometric analyses with a specialty in criterion-referenced testing.

4. Continue to score using a compensatory method. In this method, examinees must pass the exam in totality, rather than pass every component. Reliability and rater consistency have previously been found to be higher in compensatory scoring where examinees had a combined score for a battery of tests (Haladyna & Hess, 1999). In the case of the flight paramedics, this means that items from all body systems (cardiology, respiratory, etc.) should be summed up for a total score, as opposed to using a conjunctive scoring method where examinees must pass each individual component.

Exam Administration

A number of improvements were centred around portions of the administration, including oversight, handling, and distribution of the 2010 examination. The points below are based on best practice in high stakes assessment and include the practices that should be continued, as well as recommended new practices.

1. Increase the security of the exam development, delivery, and analysis. Security of the examining process must be ensured at all phases of administration. Documents created during exam preparation must be stored in a locked location. The number
of people who have access to view the exam must be kept at a minimum.

Transportation of documents must be kept to a minimum.

2. During the examination sitting, examinees must be proctored and should not have access to communication devices such as cellular telephones, cameras, or devices that can access the internet, in order to decrease chances of cheating.

3. Conduct the exam using a proctored, computer-based testing (CBT) system to reduce the possibilities of transcription error, grading errors, cheating, and to increase the security of the exam. Using a paper and pencil test for administration of the examination in 2010 and a single rater slowed down the subsequent analyses of the items and the exam. To ensure the accurate recording of examinees’ responses, the use of computer-based testing will reduce error arising from transcriptions, and will allow for a faster test analysis and reporting system. Constructed response items can still be included on a computer-based system; however, scoring will not be automated and will need to be performed by one or more raters. As well, a computer-based system is more secure than a paper system where exams can go missing.

4. Implement a manner for tracking paramedic performance from year to year to identify trends in exam scores at the paramedic level. Along with the practical Objective Structured Clinical Examinations (OSCE), this will provide an opportunity for MAC to monitor paramedic competencies over a period of time, and a variety of assessment methods.
Limitations

This study was restricted to an exploration of one recertification examination intended for a specific subset of the paramedics in the province of Ontario, Canada. While the examination process for other paramedic recertification examinations in Ontario is similar, further research will be required to determine if the findings from this study are consistent. Since there are no standardized examination processes in place for paramedic programs, it is difficult to determine the extent to which the findings of this study can inform other paramedic examination programs.

Another limitation of the study is the inability to determine whether the first attempt at using a standard setting process truly resulted in improvements. The sole comparison between the 2008 and 2010 exams was looking at passing rates. It is possible that the processes put into place in 2010 did result in improved consistency and dependability, but without strong comparisons to previous administrations this is difficult to establish. Research using several years of item and exam data would lead to more conclusive statements about the effects on consistency and dependability.

Many of the data analyses in this study were conducted on the dichotomous data only, so polytomous items were not taken into account. Future studies should incorporate both types of items in the analyses to provide a better overall picture of the whole exam.

A further limitation of the current study is my own involvement in the testing process, and the interview with a single Medical Director. The nature of the working relationship between the Medical Director and myself may have affected open and honest
dialogue during the interview, and information gathered from only one person provides a narrow viewpoint. Lastly, due to my position with the organization that conducts the examination, I had to ensure that all of the examination data I received for the study were anonymous. Hence I was not able to contact the examinees to examine their experiences and thoughts regarding the examination. Future research with this qualitative feedback could focus on examinees’ evaluation of the process.

**Conclusion**

When this study was proposed, it was thought that the improved methods of item and exam development would produce an exam that would result in more dependable and consistent classifications of paramedics seeking recertification. The study arose out of the inconsistent classification results across the previous administrations, even though it was largely the same cohort of paramedics being examined each year. The improved processes used in 2010 have likely provided a more dependable process for subsequent administrations of the examination. Nonetheless, my analyses have also identified areas for further improvement. As an example, more attention to item quality is required to produce an examination that improves dependability and increases decision consistency around the cut-score as much as possible. The findings and recommendations from this study have the potential to improve this important recertification examination. Subsequent analyses and research will provide further evidence regarding the psychometric properties of the examination and the viability of improved item writing procedures and standard setting processes. Similarly, it will be possible to explore
alternative administration procedures that meet the demands of test security and accuracy without excessive administration costs. Many research opportunities exist; this study only opens the door to the vast amount of research required to fully understand this important examination and ensure it positively supports the paramedic profession.
References


Annual Meeting of the National Council on Measurement in Education, San Diego, CA.


Appendix A

Letter of Information

“Dependability and Decision Consistency in a High-Stakes Exam for Paramedics”

Dear Medical Director,

You are being asked to participate in a study being conducted by me, Ingrid de Vries, under the supervision of Dr. Don Klinger, in the Department of Education at Queen’s University in Kingston, Ontario.

*This study has been granted clearance by the General Research Ethics Board for compliance with the “Tri-Council Policy Statement: Ethical Conduct of Research Involving Humans”, and Queen's policies.*

**What is this study about?** The purpose of this research is to examine the written exam delivered to paramedics in the winter of 2010 as part of their annual recertification process. I will be looking at the process that was used to develop the exam, as well as analyzing the results of the exam to determine if the exam was able to dependably differentiate the successful from the unsuccessful candidates. I would like to speak to you for an initial 2 hour interview, with the possibility of another 2 hour follow up interview where we will discuss the development process that was followed. There are no known physical, psychological, economic, or social risks associated with this study.

**Is participation voluntary?** Yes. You should not feel obliged to answer any material that you find objectionable or that makes you feel uncomfortable. You may also withdraw at any time with no effect on any contractual or other relationships with the investigator. If you withdraw from the study, submit a request in writing to the experimenter and supervisor specifying whether you would like to remove all or part of your data, and which part(s).

**What will happen to the responses?** We will keep your responses confidential. Only the experimenter and supervisor will have access to this information. To help us ensure confidentiality, please do not put your name on any of the research study answer sheets. The data may also be published in professional journals or presented at scientific conferences, but any such presentations will be of general findings. However, given that you are the current Medical Director of the education department at Ornge, and your name may readily be deduced, you will have the choice of your real name or a pseudonym used in the final report. Should you be interested, you are entitled to a copy of the findings. In accordance with Queen’s policy,
data will be retained for a minimum of five years; after that time, electronic data will be deleted and paper files will be shredded. Any data used for secondary analysis will contain no identifying information.

**Will there be compensation for participation?** No, there will be no compensation for your time.

**What if there are concerns?** Any questions about study participation may be directed to me, Ingrid de Vries at 7vid@queensu.ca, or my supervisor Dr. Don Klinger at don.klinger@queensu.ca or 613-533-3028. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at 613-533-6081 or chair.GREB@queensu.ca.

Sincerely,

Ingrid de Vries
Appendix B

Consent Form

“Dependability and Decision Consistency in a High-Stakes Exam for Paramedics”

Name (please print clearly): _______________________________________

1. I have read and retained the Letter of Information and have had any questions answered to my satisfaction.

2. I understand that I will be participating in the study called “Dependability and Decision Consistency in a High-Stakes Exam for Paramedics”. I understand that this means that I will be asked to discuss and explain the exam development process that was followed for the 2010 written recertification exam.

   I understand that the purpose of the study is to examine the written exam delivered to paramedics in the winter of 2010 as part of their annual recertification process and the interview will be audiotaped.

3. I understand that my participation in this study is voluntary and I may withdraw at any time. I am expected to participate in two interviews, at two hours each, for an approximate total of 4 hours. If I withdraw from the study, I may remove all or some of my data.

   I understand that every effort will be made to maintain the confidentiality of the data now and in the future. Only the experimenter and supervisor named in the Letter of Information will have access to this information. The data may also be published in professional journals or presented at scientific conferences, but any such presentations will be of general findings and will never breach individual confidentiality. Should you be interested, you are entitled to a copy of the findings.

4. Any questions about study participation may be directed to Ingrid de Vries at 7vid@queensu.ca, or the project supervisor, Dr. Don Klinger (613-533-3028), don.klinger@queensu.ca. Any ethical concerns about the study may be directed to the Chair of the General Research Ethics Board at 613-533-6081 or chair.GREB@queensu.ca.

I have read the above statements and freely consent to participate in this research.
Please sign one copy of this Consent Form and return to Ingrid de Vries. Retain the second copy for your records.

Signature: _________________________________ Date: __________________________

I would like a copy of the study results sent to me at one of the following addresses:

☐ Email address:

☐ Postal Address (please include city and postal code):
## Appendix C
### Exam Objectives

<table>
<thead>
<tr>
<th>NOCP number</th>
<th>Objectives for Critical Care exam</th>
<th>Question #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Assessment and Diagnostics</strong></td>
<td></td>
</tr>
<tr>
<td>4.5.a</td>
<td>• Evaluate oximetry waveforms</td>
<td>Not covered</td>
</tr>
<tr>
<td>4.5.b</td>
<td>• Evaluate capnographic waveforms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Explain the relationship of end-tidal carbon dioxide to arterial blood gas measurement of partial pressure of arterial carbon dioxide</td>
<td>6</td>
</tr>
<tr>
<td>4.5.g</td>
<td>• Differentiate between core and peripheral temperature monitoring</td>
<td>Not covered</td>
</tr>
<tr>
<td></td>
<td>• Explain the indications and rationale for measuring core body temperature</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Explain various means of measuring core body temperature</td>
<td></td>
</tr>
<tr>
<td>4.5.h</td>
<td>• Define pulmonary artery catheter monitoring</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>• Identify normal pulmonary artery pressures</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>• Explain indications and rationale for use of pulmonary artery catheters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Explain assessment and management of pulmonary artery catheters</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Analyze waveforms (PA)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Identify components of a normal and abnormal PA waveform and factors that can affect them.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Explain complications of pulmonary artery catheters, and their management</td>
<td></td>
</tr>
<tr>
<td>4.5.i</td>
<td>• Define central venous pressure</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>• Identify normal central venous pressure values</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Explain indications and rationale for central venous pressure monitoring</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Analyze waveforms (CVP)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Identify components of a normal and abnormal CVP waveform and factors that can affect them.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Explain complications of central venous pressure monitoring and their management</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Identify clinical presentations and hemodynamic profile of various pathologies</td>
<td></td>
</tr>
</tbody>
</table>
| 4.5.k | - Explain indications and rationale for arterial pressure monitoring  
- Analyze waveforms (arterial)  
- Identify components of a normal and abnormal arterial waveform and factors that can affect them.  
- Describe the steps to be taken to ensure the accuracy of arterial pressure values  
- Explain complications of arterial line monitoring and their management | 4  
| 4.5.k | - Identify examples of common laboratory and radiological data  
- Identify acid – base disturbances  
- Explain compensatory mechanisms in which the body responds to acid-base disturbances  
- Analyze common laboratory values including hematological and biochemical data  
- Explain the relevance of common radiological and laboratory studies to patient presentation  
- Perform a comprehensive approach to interpreting chest x-ray  
- Differentiate between solitary and disseminated lesions in the lung, air-space and interstitial disease, lobular consolidation patterns | 1,5,9,10,11,42,63,72  
| 4.5.1 | - Analyze cardiac rhythms in a 3 lead ECG | 3  
| 4.5.m | - Describe the steps involved in interpreting 12 lead ECGs, and ECGs obtained with additional leads  
- Analyze 12 and 15 lead ECGs | 21,47,48,49,56,70  
| 4.3.e | - Explain the pathophysiology of specific cardiovascular illnesses and injuries  
- Evaluate findings related to the etiology, pathophysiology, and manifestations of cardiovascular illnesses and injuries  
- CVS/hematological anatomy, physiology | 2,3  
| 4.3.d | - Explain the pathophysiology of specific neurological illnesses and injuries  
- Evaluate findings related to the etiology, pathophysiology, and manifestations of neurological illnesses and injuries | 31,43  
| 4.3.e | - Explain the pathophysiology of specific respiratory illnesses and injuries  
- Evaluate findings related to the etiology, pathophysiology, and manifestations of respiratory illnesses and injuries | 12  

82
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.f</td>
<td>Obstetrics</td>
</tr>
<tr>
<td></td>
<td>Explain the pathophysiology of specific obstetric illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Evaluate findings related to the etiology, pathophysiology, and manifestations of obstetric illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Perform multisystem assessment with specific focus on the obstetric patient</td>
</tr>
<tr>
<td>4.3.g</td>
<td>GI</td>
</tr>
<tr>
<td></td>
<td>Explain the pathophysiology of specific gastrointestinal illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Evaluate findings related to the etiology, pathophysiology, and manifestations of gastrointestinal illnesses and injuries</td>
</tr>
<tr>
<td>4.3.j</td>
<td>Trauma</td>
</tr>
<tr>
<td></td>
<td>Explain the pathophysiology of specific musculoskeletal illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Evaluate findings related to the etiology, pathophysiology, and manifestations of musculoskeletal illnesses and injuries</td>
</tr>
<tr>
<td>4.3.k</td>
<td>Immune system</td>
</tr>
<tr>
<td></td>
<td>Explain the pathophysiology of specific immune system illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Evaluate findings related to the etiology, pathophysiology, and manifestations of immune system illnesses and injuries</td>
</tr>
<tr>
<td>4.3.l</td>
<td>Endocrine</td>
</tr>
<tr>
<td></td>
<td>Explain the pathophysiology of specific endocrine illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Evaluate findings related to the etiology, pathophysiology, and manifestations of endocrine illnesses and injuries</td>
</tr>
<tr>
<td>4.3.o</td>
<td>Neonatal</td>
</tr>
<tr>
<td></td>
<td>Explain the pathophysiology of specific neonatal illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Evaluate findings related to the etiology, pathophysiology, and manifestations of neonatal illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Perform multisystem assessment with specific focus on the neonatal patient</td>
</tr>
<tr>
<td>Pediatrics</td>
<td>Explain the pathophysiology of specific pediatric illnesses and injuries</td>
</tr>
<tr>
<td></td>
<td>Evaluate findings related to the etiology, pathophysiology, and manifestations of pediatric illnesses and injuries</td>
</tr>
</tbody>
</table>
Appendix D
Interview Questions

1. How many paramedics do you think should be passing the exam? Is there a goal?
2. What is the item difficulty that you are targeting?
3. Do you or the Medical Advisory Committee (MAC) have predefined thoughts on: pass rate, test length, discrimination, or using an item bank?
4. How many people is it realistic to have on a standard setting panel?
5. Do you feel the exam evaluated the areas that it was intended to measure?
6. What effect(s) do you think the development process had on the results?
7. Was sufficient time spent on the development process?
8. How sure would you like to be that the classification is correct?
9. What factors would you be willing to compromise on – number of items? Number of judges on the panel?
10. For future exams, what would you like to see done differently? What would you like to remain the same?
11. If we began piloting items, who would take the pilot exam? How many items would be reasonable to pilot?
12. Who do feel is best suited to write the items – Physicians? Medical Directors? Paramedics?
13. What would be the impact of reverting to the previous exam system?
### Appendix E

**Analysis of Dichotomous Items**

<table>
<thead>
<tr>
<th>Item</th>
<th>B Index</th>
<th>Agreement</th>
<th>Item Φ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.11</td>
<td>0.56</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.16</td>
<td>0.52</td>
<td>0.15</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.16</td>
<td>0.65</td>
<td>0.15</td>
</tr>
<tr>
<td>5</td>
<td>0.24</td>
<td>0.79</td>
<td>0.37</td>
</tr>
<tr>
<td>6</td>
<td>0.05</td>
<td>0.7</td>
<td>0.08</td>
</tr>
<tr>
<td>7</td>
<td>0.2</td>
<td>0.73</td>
<td>0.24</td>
</tr>
<tr>
<td>9</td>
<td>0.09</td>
<td>0.69</td>
<td>0.15</td>
</tr>
<tr>
<td>10</td>
<td>0.07</td>
<td>0.66</td>
<td>0.11</td>
</tr>
<tr>
<td>13</td>
<td>0.29</td>
<td>0.72</td>
<td>0.29</td>
</tr>
<tr>
<td>14</td>
<td>0.04</td>
<td>0.32</td>
<td>0.07</td>
</tr>
<tr>
<td>17</td>
<td>0.05</td>
<td>0.73</td>
<td>0.11</td>
</tr>
<tr>
<td>18</td>
<td>-0.01</td>
<td>0.43</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>0.06</td>
<td>0.76</td>
<td>0.26</td>
</tr>
<tr>
<td>21</td>
<td>0.08</td>
<td>0.35</td>
<td>0.11</td>
</tr>
<tr>
<td>22</td>
<td>-0.06</td>
<td>0.7</td>
<td>-0.09</td>
</tr>
</tbody>
</table>

85
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>0.26</td>
<td>0.77</td>
<td>0.34</td>
</tr>
<tr>
<td>24</td>
<td>0.15</td>
<td>0.53</td>
<td>0.14</td>
</tr>
<tr>
<td>25</td>
<td>0.18</td>
<td>0.77</td>
<td>0.31</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>-0.03</td>
<td>0.35</td>
<td>-0.03</td>
</tr>
<tr>
<td>28</td>
<td>0.16</td>
<td>0.52</td>
<td>0.15</td>
</tr>
<tr>
<td>29</td>
<td>0.09</td>
<td>0.52</td>
<td>0.09</td>
</tr>
<tr>
<td>31</td>
<td>0.24</td>
<td>0.2</td>
<td>0.22</td>
</tr>
<tr>
<td>32</td>
<td>0.15</td>
<td>0.6</td>
<td>0.15</td>
</tr>
<tr>
<td>33</td>
<td>0.22</td>
<td>0.5</td>
<td>0.21</td>
</tr>
<tr>
<td>34</td>
<td>-0.14</td>
<td>0.37</td>
<td>-0.12</td>
</tr>
<tr>
<td>36</td>
<td>0.15</td>
<td>0.59</td>
<td>0.14</td>
</tr>
<tr>
<td>38</td>
<td>0.25</td>
<td>0.75</td>
<td>0.29</td>
</tr>
<tr>
<td>39</td>
<td>0.41</td>
<td>0.7</td>
<td>0.37</td>
</tr>
<tr>
<td>41</td>
<td>0.16</td>
<td>0.72</td>
<td>0.21</td>
</tr>
<tr>
<td>42</td>
<td>0.32</td>
<td>0.69</td>
<td>0.3</td>
</tr>
<tr>
<td>43</td>
<td>0.39</td>
<td>0.66</td>
<td>0.35</td>
</tr>
<tr>
<td>44</td>
<td>-0.02</td>
<td>0.73</td>
<td>-0.02</td>
</tr>
<tr>
<td>45</td>
<td>0.1</td>
<td>0.36</td>
<td>0.13</td>
</tr>
<tr>
<td>47</td>
<td>0</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

86
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>0.02</td>
<td>0.72</td>
<td>0.07</td>
</tr>
<tr>
<td>49</td>
<td>0.19</td>
<td>0.63</td>
<td>0.18</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.7</td>
<td>0.04</td>
</tr>
<tr>
<td>53</td>
<td>-0.05</td>
<td>0.55</td>
<td>-0.04</td>
</tr>
<tr>
<td>56</td>
<td>0.26</td>
<td>0.53</td>
<td>0.24</td>
</tr>
<tr>
<td>57</td>
<td>0.33</td>
<td>0.56</td>
<td>0.31</td>
</tr>
<tr>
<td>62</td>
<td>0.02</td>
<td>0.55</td>
<td>0.03</td>
</tr>
<tr>
<td>63</td>
<td>0.25</td>
<td>0.72</td>
<td>0.26</td>
</tr>
<tr>
<td>64</td>
<td>-0.02</td>
<td>0.67</td>
<td>-0.01</td>
</tr>
<tr>
<td>67</td>
<td>0.15</td>
<td>0.59</td>
<td>0.14</td>
</tr>
<tr>
<td>68</td>
<td>0.01</td>
<td>0.7</td>
<td>0.04</td>
</tr>
<tr>
<td>69</td>
<td>0.09</td>
<td>0.73</td>
<td>0.15</td>
</tr>
<tr>
<td>70</td>
<td>0.15</td>
<td>0.59</td>
<td>0.14</td>
</tr>
<tr>
<td>71</td>
<td>-0.02</td>
<td>0.67</td>
<td>-0.01</td>
</tr>
<tr>
<td>72</td>
<td>0.1</td>
<td>0.66</td>
<td>0.11</td>
</tr>
<tr>
<td>73</td>
<td>0.12</td>
<td>0.67</td>
<td>0.14</td>
</tr>
<tr>
<td>74</td>
<td>-0.07</td>
<td>0.15</td>
<td>-0.06</td>
</tr>
<tr>
<td>76</td>
<td>0.09</td>
<td>0.6</td>
<td>0.12</td>
</tr>
<tr>
<td>77</td>
<td>-0.06</td>
<td>0.59</td>
<td>-0.07</td>
</tr>
</tbody>
</table>
## Appendix F
### Analysis of Discriminating Items

<table>
<thead>
<tr>
<th>Item</th>
<th>B Index</th>
<th>Agreement</th>
<th>Item $\Phi$</th>
<th>$r_{pb}$</th>
<th>$p_{\text{pred}} - p_{\text{actual}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.16</td>
<td>0.65</td>
<td>0.15</td>
<td>0.38</td>
<td>-0.13</td>
</tr>
<tr>
<td>5</td>
<td>0.24</td>
<td>0.79</td>
<td>0.37</td>
<td>0.27</td>
<td>-0.12</td>
</tr>
<tr>
<td>7</td>
<td>0.2</td>
<td>0.73</td>
<td>0.24</td>
<td>0.32</td>
<td>-0.01</td>
</tr>
<tr>
<td>13</td>
<td>0.29</td>
<td>0.72</td>
<td>0.29</td>
<td>0.39</td>
<td>0.07</td>
</tr>
<tr>
<td>18</td>
<td>-0.01</td>
<td>0.43</td>
<td>0</td>
<td>0.25</td>
<td>-0.04</td>
</tr>
<tr>
<td>23</td>
<td>0.26</td>
<td>0.77</td>
<td>0.34</td>
<td>0.31</td>
<td>-0.18</td>
</tr>
<tr>
<td>24</td>
<td>0.15</td>
<td>0.53</td>
<td>0.14</td>
<td>0.25</td>
<td>0.16</td>
</tr>
<tr>
<td>28</td>
<td>0.16</td>
<td>0.52</td>
<td>0.15</td>
<td>0.48</td>
<td>0.35</td>
</tr>
<tr>
<td>31</td>
<td>0.24</td>
<td>0.2</td>
<td>0.22</td>
<td>0.41</td>
<td>0.32</td>
</tr>
<tr>
<td>33</td>
<td>0.22</td>
<td>0.5</td>
<td>0.21</td>
<td>0.29</td>
<td>0.27</td>
</tr>
<tr>
<td>39</td>
<td>0.41</td>
<td>0.7</td>
<td>0.37</td>
<td>0.47</td>
<td>0.18</td>
</tr>
<tr>
<td>42</td>
<td>0.32</td>
<td>0.69</td>
<td>0.3</td>
<td>0.41</td>
<td>0.24</td>
</tr>
<tr>
<td>43</td>
<td>0.39</td>
<td>0.66</td>
<td>0.35</td>
<td>0.46</td>
<td>0.29</td>
</tr>
<tr>
<td>49</td>
<td>0.19</td>
<td>0.63</td>
<td>0.18</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>56</td>
<td>0.26</td>
<td>0.53</td>
<td>0.24</td>
<td>0.43</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>0.33</td>
<td>0.56</td>
<td>0.31</td>
<td>0.46</td>
<td>0.14</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>57</td>
<td>0.25</td>
<td>0.72</td>
<td>0.26</td>
<td>0.33</td>
<td>0.11</td>
</tr>
<tr>
<td>63</td>
<td>0.09</td>
<td>0.73</td>
<td>0.15</td>
<td>0.26</td>
<td>-0.1</td>
</tr>
<tr>
<td>69</td>
<td>0.15</td>
<td>0.59</td>
<td>0.14</td>
<td>0.26</td>
<td>0.41</td>
</tr>
<tr>
<td>70</td>
<td>0.1</td>
<td>0.66</td>
<td>0.11</td>
<td>0.28</td>
<td>0.23</td>
</tr>
<tr>
<td>72</td>
<td>0.12</td>
<td>0.67</td>
<td>0.14</td>
<td>0.33</td>
<td>0.21</td>
</tr>
</tbody>
</table>
## Appendix G
### Analysis of Problematic Items

<table>
<thead>
<tr>
<th>Item</th>
<th>B index</th>
<th>Agreement</th>
<th>Item Φ</th>
<th>$r_{pb}$</th>
<th>$p_{pred} - p_{actual}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>.09</td>
<td>.69</td>
<td>.15</td>
<td>.15</td>
<td>-.05</td>
</tr>
<tr>
<td>10</td>
<td>0.07</td>
<td>0.66</td>
<td>0.11</td>
<td>0.12</td>
<td>-0.04</td>
</tr>
<tr>
<td>14</td>
<td>0.04</td>
<td>0.32</td>
<td>0.07</td>
<td>-0.01</td>
<td>0.61</td>
</tr>
<tr>
<td>17</td>
<td>0.05</td>
<td>0.73</td>
<td>0.11</td>
<td>0.11</td>
<td>-0.18</td>
</tr>
<tr>
<td>19</td>
<td>0.06</td>
<td>0.76</td>
<td>0.26</td>
<td>0.10</td>
<td>-0.16</td>
</tr>
<tr>
<td>21</td>
<td>0.08</td>
<td>0.35</td>
<td>0.11</td>
<td>0.05</td>
<td>0.62</td>
</tr>
<tr>
<td>27</td>
<td>-0.03</td>
<td>0.35</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.43</td>
</tr>
<tr>
<td>29</td>
<td>0.09</td>
<td>0.52</td>
<td>0.09</td>
<td>0.04</td>
<td>0.31</td>
</tr>
<tr>
<td>32</td>
<td>0.15</td>
<td>0.6</td>
<td>0.15</td>
<td>0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>34</td>
<td>-0.14</td>
<td>0.37</td>
<td>-0.12</td>
<td>0.13</td>
<td>0.39</td>
</tr>
<tr>
<td>38</td>
<td>0.25</td>
<td>0.75</td>
<td>0.29</td>
<td>0.09</td>
<td>0.15</td>
</tr>
<tr>
<td>44</td>
<td>-0.02</td>
<td>0.73</td>
<td>-0.02</td>
<td>-0.12</td>
<td>-0.24</td>
</tr>
<tr>
<td>45</td>
<td>0.1</td>
<td>0.36</td>
<td>0.13</td>
<td>0.06</td>
<td>0.47</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.7</td>
<td>0.04</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>53</td>
<td>-0.05</td>
<td>0.55</td>
<td>-0.04</td>
<td>0.13</td>
<td>-0.05</td>
</tr>
<tr>
<td>62</td>
<td>0.02</td>
<td>0.55</td>
<td>0.03</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>-0.02</td>
<td>0.67</td>
<td>-0.01</td>
<td>0.12</td>
<td>-0.07</td>
</tr>
<tr>
<td>67</td>
<td>0.15</td>
<td>0.59</td>
<td>0.14</td>
<td>0.10</td>
<td>0.29</td>
</tr>
<tr>
<td>71</td>
<td>-0.02</td>
<td>0.67</td>
<td>-0.01</td>
<td>-0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>74</td>
<td>-0.07</td>
<td>0.15</td>
<td>-0.06</td>
<td>0.02</td>
<td>0.39</td>
</tr>
<tr>
<td>76</td>
<td>0.09</td>
<td>0.6</td>
<td>0.12</td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>77</td>
<td>-0.06</td>
<td>0.59</td>
<td>-0.07</td>
<td>0.02</td>
<td>-0.07</td>
</tr>
</tbody>
</table>