2011

Examining Construct Stability Across Career Stage Cohorts

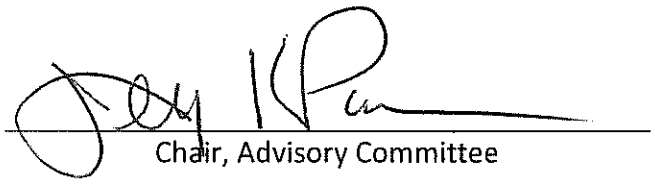
Deborah L. Kinney
Eastern Kentucky University

Examining Construct Stability Across Career Stage Cohorts

By

Deborah L. Kinney

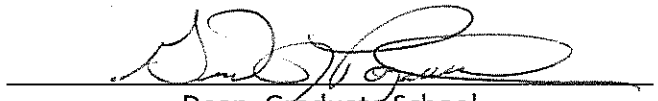
Thesis Approved:



Chair, Advisory Committee

Member, Advisory Committee

Member, Advisory Committee



Dean, Graduate School

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Masters of Science degree at Eastern Kentucky University, I agree that the Library shall make it available to borrowers under rules of the Library. Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgment of the source is made. Permission for extensive quotation from or reproduction of this thesis may be granted by my major professor, or in [his/her] absence, by the Head of Interlibrary Services when, in the opinion of either, the proposed use of the material is for scholarly purposes. Any copying or use of the material in this thesis for financial gain shall not be allowed without my written permission.

Signature Deborah L Hensley

Date 4/18/11

Examining Construct Stability Across Career Stage Cohorts

By

Deborah L. Kinney

Bachelor of Arts
University of South Florida
Tampa, FL
2003

Submitted to the Faculty of the Graduate School of
Eastern Kentucky University
in partial fulfillment of the requirements
for the degree of
MASTER OF SCIENCE
May, 2011

DEDICATION

This thesis is dedicated to my husband Michael and my children Reice, Garett, and Liadan for keeping my reality grounded during this entire process.

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Jerry Palmer, for his patience and input throughout this process. I would also like to thank my other committee members, Dr. Richard Osbaldiston and Dr, Paul Erickson, for their valuable feedback, perspective, and support. I would like to thank Dr. Tom O'Neill from the American Board of Family Medicine for the opportunity to embark on this project, and for making the entire experience possible. Finally, I would like to thank my family, for the most important learning experience, my life outside the boundaries of academia. They have taught me that the most important lessons will never be learned from a book.

ABSTRACT

The purpose of this study is to evaluate construct stability of the same certification test taken at different points in a test taker's career, taking into account changes is experience over time. A single medical certification exam administration was used to analyze the construct stability of the certification exam across testing cohorts at varied stages in their medical careers. The Rasch model was used for item analysis to calibrate the difficulty hierarchy of the exam items for each cohort. Correlations between the item difficulty hierarchies for each cohort supported the overall construct stability of the certification exam. Individual item function for each cohort was analyzed through a differential item functioning (DIF) procedure, which showed less than 5% overall DIF, again supporting the construct stability of the examination. The support for the stability of the construct measured by the exam is a necessary condition in the process establishing the validity of the exam, making the information in this study valuable for a variety of testing implications.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE PAGE

1. INTRODUCTION

The purpose of this study is to evaluate the concept of construct stability of a medical certification examination taken at points over time in a physicians career. This will be shown through establishing working definitions of constructs and stability of constructs, providing insight into how these ideas are concretely measured. A standardized medical certification exam will serve as the data source for completing a full construct stability analysis to document the process of ensuring a valid, stable construct measure in assessment.

Constructs and Establishing Construct Stability

Constructs are the means by which science orders observations. Educational and psychological constructs are generally attributes of people, situations, or treatments reflected in test performance, ratings and other observations. "We take it on faith that the universe of our observation can be ordered and subsequently explained with a comparatively small number of constructs" (Stenner, Smith, & Burdick, 1983, p. 306). The meaning of a measurement depends on a construct theory. The simple fact that numbers are assigned to observations in a systematic manner implies some hypothesis about what is being measured. Instruments (e.g., tests) are the link between theory and observation, and scores are the readings or values generated by instruments. The validity of any given construct theory is a matter of degree, dependent, in part, on how well it

predicts variation in item scale values and person scores and the breadth of these predictions (Stenner, et al., 1983).

The process of ascribing meaning to scores produced by a measurement procedure is generally recognized as the most important task in developing an educational or psychological measure, be it an achievement or performance test, interest inventory, or personality scale. This is referred to as construct validity (Cronbach, 1971; Cronbach & Meehl, 1955; Messick, 1975). Testing construct validity involves a family of methods and procedures to confirm that the test measures a trait or theoretical construct (Stenner, et al., 1983). In order to demonstrate that a measurement tool has construct validity, it must measure the same latent trait across all test administration. A measurement tool must report the same value for the same amount of a construct regardless of the sample and point in time of the test administration. A test does not have construct validity if it does not measure what it is designed to measure and does not have consistent stability over time (Cook & Campbell, 1979).

Construct Stability in Testing Over Time

The changes in knowledge and ability over time can have a significant effect on the construct stability of a performance measure. The stability of a measure is established by evaluating the individual test items to determine whether they are stable indicators of the latent trait they have been designed to measure. The items must maintain consistency regardless of the respondent sample or test administration. When this consistency or invariance is maintained, an item is considered to have acceptable construct stability. When all of the items on a given test are found to function in a way

that every respondent with comparable ability levels has an equal chance of answering the items correctly regardless of item difficulty, construct stability is present, therefore construct validity for the measure is supported.

The way in which constructs are expressed can be described in terms of a hierarchy of difficulty spanning from easy to hard. The difficulty order of items in a measure should remain consistent regardless of the testing sample, because that hierarchy defines the construct. When the basic hierarchical structure of the item difficulty remains the same across different cohorts of examinees, such as in recent graduates and seasoned professionals, the measure shows strong construct stability.

Physicians are very likely to have an array of different learning experiences and professional opportunities over the course of their medical careers. There are possibilities that different performance on the constructs over time is related to changes in the cohorts over time. Consider a cohort of physicians who take an initial certification exam immediately after finishing residency training. There may be a noted difference in the cohort's performance on the exam compared to the performance of a cohort that is perhaps 6 or 12 years in the field since initial certification. Some things that affect measurement of these groups' knowledge and performance include the recency of study of information and mastery of the material. The latter could promote better performance on the more theoretical items and poorer performance on the more applied areas of the measure. This effect may also be indeed converse for those who have been in an applied practice position for some time. This experience may boost performance on applied test items and decrease knowledge performance. Logically, if practitioners do not remain up-to-date with current practices, their performance on a standard certification exam may

theoretically be a strong indicator of these deficits. In this way, new medical program graduates are thought to have more recent access to the most current knowledge and techniques, therefore resulting in stronger certification exam performance. It is a good assumption that recent graduates are outscore physicians who have been certified longer on board examinations. As physicians progress through their career, their practice may become more specialized toward a particular patient population, which might make them less proficient in broad-spectrum practice and recertification testing, as is discussed here.

For the current study, breaking the examinees into cohorts by initial and recertification cycles provides an opportunity to further analyze the functioning of each item and examine the response patterns for each construct within each group. The chosen model of analysis for this purpose is the Rasch model of Item Response Theory (IRT) and Differential Item Functioning (DIF).

Item Response Theory and Rasch Model in Testing

IRT is a family of psychometric models that adjust for the ability of the candidates when estimating the difficulty of items and for the difficulty of the items when estimating the ability of the candidates. IRT comprises a collection of modeling techniques for the analysis of item-level data obtained to measure variation among persons. (Edelen, Thissen, Teresi, Kleinman, & Ocepeck-Welikson, 2006).

A central feature of the Rasch based IRT model (described more fully below) is a table of expected probabilities designed to address the key question: When a person with this ability (number of test items correct) encounters an item of this difficulty (number of persons who succeeded on the item), what is the likelihood that this person gets this item

correct? The answer to this question is that the probability of success depends on the difference between the ability of the person and the difficulty of the item. Wright & Panchepakesan (1969) in their work on sample-free and item-free measurement describes the procedure for sample-free item analysis as based on a very simple model designed by Rasch (1960) for what occurs in testing each time an individual encounters a test item. The Rasch model proposes that the outcome of the interaction between the individual and the test item is governed by the ability of the person and difficulty of the item itself, nothing more (Wright & Panchepakesan, 1969).

The Rasch model then allows for test item calibrations. Calibrating items gives each item a place in the difficulty continuum for that specific time of measurement, based on the performance of the test takers at that time. The place on the continuum corresponds with the proportion of correct responses for the current testing cohort. The smaller the proportion of correct responses to a specific item, the higher the difficulty of an item and hence the higher the item's location on the scale. The higher the calibration of an item, the more difficult the item was for the current cohort. These calibrations translate into the ability estimates for the entire cohort on each individual item. The placement of the item difficulty for each cohort allows for comparisons to be made between different groups of test takers to show similarities in item performance, based on the calibration scores of the individual items. Thus, we might say that two instruments are well calibrated psychometrically because the items are matched according to item difficulty or item discrimination. The correlations between these calibrations allow us to make inferences into how strongly related the items are for different groups across time. The calibration of measuring instruments must be independent of the test items that

happen to be used for calibration, and the measurement of test items must be independent of the instrument that happens to be used for measurement. This form of object-free instrument calibration and instrument-free object calibration make possible the conditions for generalization of measurement beyond the specific instrument or test used, enabling comparisons of objects on similar but not identical instruments (Wright, 1977).

In testing, the calibration of test ease or difficulty must be independent of the individuals' scores being used for the calibration. The measurement of person ability must be independent of the individual test items being used for measuring that ability. Ideally, when we compare individual items against each other in the process of test calibration, it should not matter which individual responses are used for the comparison. This type of test/item calibration allows for the construction of tests that have uniform meaning regardless of the groups that are chosen for measurement by them (Wright, 1977). The Rasch model also incorporates a method for ordering persons (e.g., from a sample of school children) according to their ability and ordering items (e.g., from a diagnostic test of numerical computations) according to their difficulty (Bond & Fox, 2007). The 1-parameter logistic (1PL) model for dichotomous data, also known as the dichotomous Rasch model, is:

$$\Pr\{X_{ni} = 1\} = \frac{e^{\beta n - \delta_i}}{1 + e^{\beta n - \delta_i}}$$

Where:

$Pr\{X_{ni} = 1\}$ = is the probability of a correct response by person *n* on item *i*

β_n is the ability of person *n*
δ_i is the difficulty of item *i*
e is the base of the natural logarithm

According to the Program Committee of the Institute for Objective Measurement (2000), "Objective measurement is the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured," (http://www.rasch.org/define.htm). An assumption of the Rasch model is that across persons measured, different brands of instruments, and instrument users, an objective measurement remains constant (Linacre, 2008). Objective measurement produces a reference standard in place of a quantitative value, thereby generating uniform terms so research relevant to a particular variable can be conducted and compared.

It is necessary that definitions relevant to all aspects of the Rasch model are clearly understood before continuing with the use of the model. The Institute of Objective Measurement (IOM, 2000) developed a comprehensive list of these definitions. The Rasch model utilizes a single standard metric to measure person ability and item difficulty. Ability is defined as a person's level of success on the measurement evaluating a variable, such as how many questions a respondent answered correctly on an exam. Difficulty is a measure of the level of resistance to success an item has. In other words, difficulty is how challenging an item is for respondents. An example of a low level of difficulty would be an item that most respondents answer correctly, whereas an example of a high level of difficulty would be an item that none or few respondents answer correctly. This difficulty measure is mathematically converted to logits, which are defined by IOM as the units of measure used in the Rasch model for calibrating items and measuring persons. Mathematically speaking, a logit is a log odds transformation of

the probability of a correct response. The result is the probability, in logits, that a respondent will answer an item correctly.

2. LITERATURE REVIEW

Previous Rasch Research

There are a wide variety of important applications in the fields of testing and measurement. Rasch measurement is the stochastic standard of measurement for monotonically increasing variables (American Board of Family Medicine, ABFM, 2009). Monotonically increasing variables are those in which a larger raw score indicates a larger amount of the construct being measured. The utility of the Rasch model allows for a wider variety of practical applications, reaching a number of diverse professional fields. This information can be used to set individual goals for employees. The diversity of the Rasch model further allows evaluators to make comparisons from employees to one another, or to assess individual improvement or decreased ability over time. Group level analysis across performance dimensions can be additionally conducted after Rasch item calibrations have been completed. These group comparisons may discover that a certain group or cohort may excel in one area while needing improvement in others**.** Improvement goals can then be set relative to the expected group performance at that specific point in time. There are several other applications for the Rasch model in areas outside of specific measurement.

One of the uses of the Rasch model outside of the specific field measurement can be to analyze selection information which aides in hiring decisions. The ability to calibrate items by difficulty and separately calibrate persons by ability make the model a good fit for this task**.** By allowing job applicants to be placed on an ability based scale,

hiring decisions can be easily based on performance because the applicants at the top of the scale have the highest ability and those at the bottom have the lowest ability.

Much like the procedures of assessing performance ability in selection processes, other professions require certification or licensure for entry into practice**.** These tests are designed specifically to measure the same construct over time, so the construct should be stable across test administrations. The purpose of this study is to evaluate construct stability of the same certification test taken at different points in a test taker's career, taking into account changes in experience over time.

In previous research, testing organizations use classical test theory or a regression approach as selection models or benchmarks to assess knowledge in areas such as education and certification (Hollman, 1973). Simply put, when any group of people take a test, it is assumed that regression leads to accurate selection by predicting success or future ability (Cole, 1993; Darlington, 1971; Einhorn & Bass, 1971; Thorndike, 1971; as cited in Wright, Mead, & Draba, 1976). Differently, the Rasch model equates the probability of a success outcome from person ability as well as item difficulty and develops measures of relative ability and difficulty on an exam (IOM, 2000; Wright et al., 1976). The choice of the Rasch model over that of a regression approach allows for test scores compared over time to be more useful. This utility is due to the separation in the measurement of person ability and item difficulty, not only individual performance.

Rasch Modeling and Certification Exams

In the medical field, it is important for physicians to be professionally certified. In order to earn certification, physicians must meet several criteria including passing a certification examination. The ABFM is one organization that provides such certification exams. The primary construct being measured on the ABFM exam can be closely tied to cognitive knowledge and practice related problem-solving ability relevant to family medicine. This knowledge of family medicine that physicians are required to possess is always in a state of change. Due to the constant flux of the required knowledge, it is highly unlikely that the construct is completely stable over time. With this in mind, it is important that construct stability is not measured only in longitudinal terms, but measured between physicians taking the exam at various stages in their certification careers.

One of the primary requirements for Rasch measurement is that there is invariance in the comparisons among both the test items and the test takers. If two items are presented to an examinee, the easier of the two items will always have the higher probability of being answered correctly, regardless of the ability level of each examinee. This requirement makes this method of analysis ideal for the current study, where individual exam performance on individual test items is of high importance.

Differential Item Functioning and Construct Validity

In order to demonstrate that a test is valid, each of the individual items should be truly unbiased. This absence of bias allows each person of equal ability to have an equal chance of answering items correctly**.** Unbiased items do not function differently across

different groups of respondents. Each item should not provide an advantage or disadvantage to examinees based on group membership. When an item is biased, it favors one group of test takers over another.

One way to test for item bias is through an analysis of DIF. DIF occurs when members of different groups exhibit a significant difference in the probability of success on an item (Cole, 1993, Roever, 2005). Cohorts from differing points in time can be experimentally stratified or statistically adjusted to be equal on the latent trait measured by a test and should therefore be equally likely to answer the item correctly, therefore the item differences can be attributed to the functioning of the item, not the cohort.

By detecting potentially biased items, DIF analyses can be used for demonstrating construct stability/invariance across cohorts and performance over time. This opens the possibility that biased items contain extraneous factors of difficulty and knowledge which are unrelated to the construct being measured by the test (Williams, 1997; Zumbo, 1999). It is vital to note that DIF does not identify items as definitively biased, but it does indicate those with potential to be biased. In these terms, DIF becomes a necessary, but not sufficient condition for item bias (Perrone, 2006). Items showing evidence of possible bias are selected out for further detailed analysis. These further investigations may reveal not only item bias, but possible group differences in ability.

Rasch based DIF is used for detecting bias in test items, examining the differences in item difficulty after correcting for test taker ability (Angoff, 1993). The response to an item is a direct function of the ability level of the examinee and the characteristics of the item (Embretson & Reise, 2000). This then results in the use of the Rasch model equating to a repeated measures t-test evaluating the difficulty of an item for one testing

group compared to another. The t-test shows the differences in mean item difficulty for each cohort, over the joint standard error of both cohorts. This illustrates that for each item the evaluation is a t-test comparing the difficulty parameter of the focal group with the reference group.

The results of the t-test series are often best illustrated as a scatter plot: the calibration of item difficulty on the x-axis for one subgroup, and the calibration of difficulty for the same item on the y-axis for the comparison subgroup (See Figure 1).The 95% confidence interval lines are added to further clarify the presence of DIF on the graph. These confidence intervals clearly indicate with 95% confidence that those test items that lie inside the lines do not show evidence of DIF. Those items that are outside of the confidence interval lines are thought to function differently and are investigated further for potential bias. The items that are found to be significantly different need to be viewed by test developers and subject matter experts on the relevant test material to accurately identify the potential causes of this difference.

The effective evaluation and resolution of items to ensure that there is no item bias and the resulting construct stability using the Rasch model is especially important in high-stakes testing, such as in academics. Certification examinations rely on the DIF analysis to specifically identify items that may need to be revaluated for bias for or against a particular testing group.

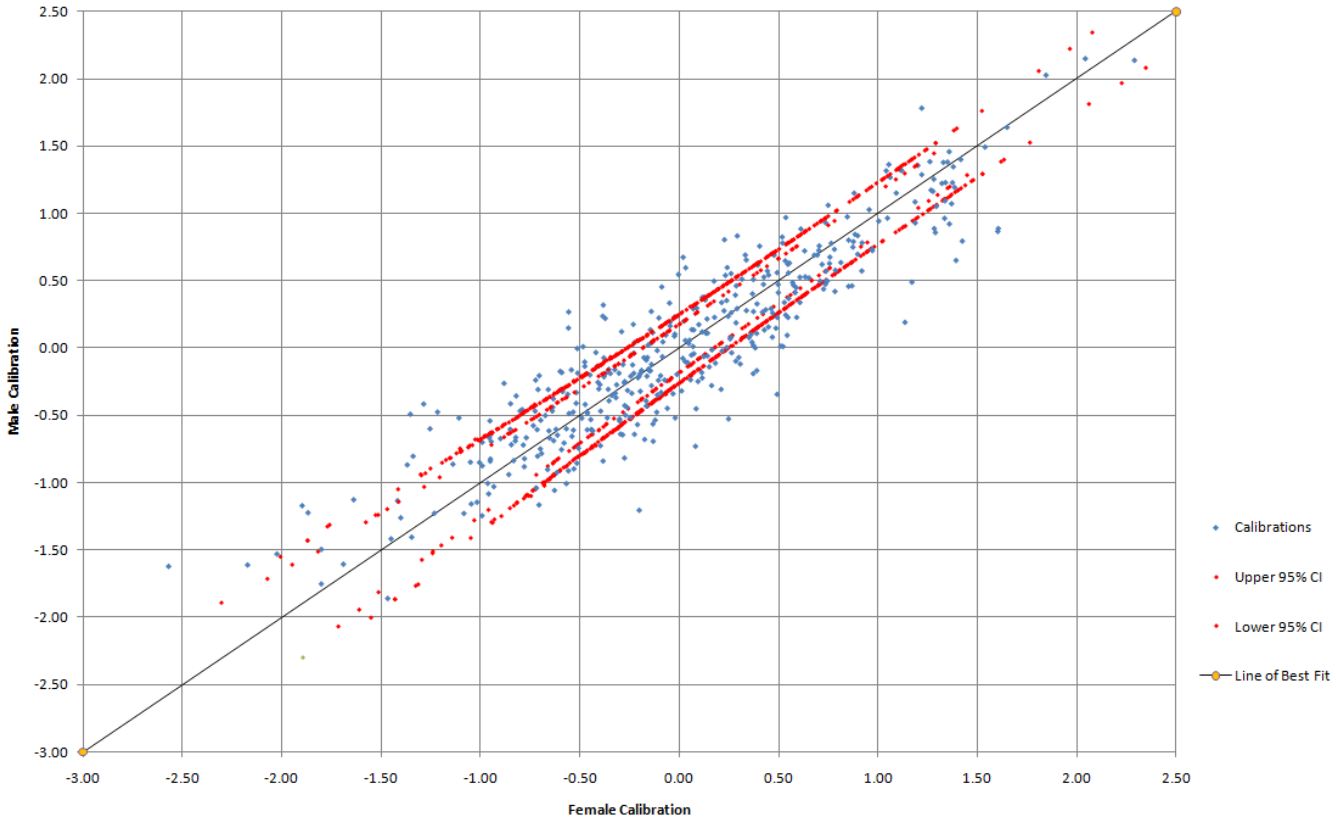


Figure 1. Sample Calibration Scatter Plot

Scatter plot showing DIF of female and male calibrations with 95% confidence intervals (fictional data).Points outside the red bands fall outside of the 95% confidence interval and must be evaluated for bias.

3. RATIONALE and HYPOTHESES

At the ABFM, the certification examination is given for initial certification or as a requirement to renew or maintain certification. Under current maintenance of certification physicians must recertify every 7 years, with eligibility to take the recertification exam in the 6th year. An interesting question to explore is if the constructs on the initial certification test and the recertification test are equivalent within stochastic limits. As discussed earlier, the ability of the physician taking these exams may vary due to individual differences at stages in the physician's career. That is, at one point in a physician's career, a test may have evidence of construct validity. Yet, from the time construct validity is initially established to a point that it is administered a second time, the construct being measured may appear different. These differences in construct will exist at a more individualized level as opposed to variations in the global construct itself. The change over time may be due to changes in practice standards or in individual material grasp and preparation. In the present study, the cohorts of physicians taking the certification exam for initial certification and renewal of the certification throughout cycles over time will be compared as a means of measuring construct stability.

The ABFM asserts that because the same certificate is issued to physicians who took the exam for initial certification and those who took it for recertification, the same construct is intended to be at work. With this assertion in mind, it becomes important to consistently apply the same test and the same passing standard to create stability. Consequently, calibrations from the initial certification cohort and the recertification

cohorts should be positively correlated. Therefore, the following hypothesis will be tested:

Hypothesis 1: There will be a significant positive correlation between the item calibrations based upon the initial certification cohort each recertification cohort.

Based on the assumption that the construct is the same for each item on the certification exam within cohorts taking it for initial certification and recertification, the hierarchy of item difficulty should also be the same regardless of the ability level of any of the cohorts. That is, if the two groups are of comparable ability and each item demonstrates construct stability, the pattern of item difficulty should be the same. Correlating the individual item calibrations for each cohort allows for the relationship between each item for each cohort to be examined. The presence of this significant positive correlation will allow us to maintain the assumptions of construct stability for the base exam regardless of cohort performance.

Additionally from the previous research that there are differences in knowledge base over the course of a physician's career, it is expected that sets of items will show small amounts of DIF for different cohort cycles.

Hypothesis 2: There will be DIF found in less than 5% of the sets of items compared across cohorts, based on each cohorts' cycle away from initial certification.

Based again on the assumptions of this study that the ABFM certification exam exhibits construct stability across time and cohort, this hypothesis will be supported by the low percentage of DIF found. Higher levels of DIF between the cohorts would show possible bias in the question structure for some of the cohorts taking the examination. This type

of variance across the cohorts would be detrimental to the overall construct stability of the exam and require attention to stabilize the items. The parameter of 5% is set for this study due to the fact that in larger sample sizes there is a possibility for higher levels of spurious DIF evidence. Keeping this parameter low addresses this possibility while maintaining the ability to verify construct stability.

4. METHOD

Participants

The ABFM offers two versions of the primary certification examination within a given year. Examination versions are randomly assigned to candidates unless it is being taken in the pencil and paper format or they already have taken one version, in which case they are assigned the other version. This study will use archival data from the ABFM 2009 summer Primary Certification/Recertification exam.

The ABFM summer 2009 primary certification/recertification examination (MC_FPE) was completed by 10,801 candidates. Of these examinees, there were 2440 initial certification candidates and 8361 recertification candidates. The results for initial certification candidates, which are the primary measure of comparison in the current study, were similar to previous initial certification cohorts. One hundred nine were offered the opportunity to invalidate the test results because of problems with the administration of their examination (ABFM, 2009).

For the purposes of this study, the candidates will be grouped according to certification cycle cohorts, following the maintenance of certification 7 year recertification plan guidelines. The cohorts of interest will be the initial certifications for 2009, one certification cycle out (2002-2003), two recertification cycles out (1995-1997), and three recertification cycles out (1988-1991). The dates of these cohorts grow as they increase in numbers of certifications back to account for the ability of the examinee's to take the exam in their 6^{th} or 7^{th} year of certification. The cohorts from 1988-2002 generally represent diplomates that successfully maintained their certification since their

initial certification. Surprisingly, almost no candidates certified in 2003 chose to take the examination in the 6th year, which is 2009. Demographic information about the test-takers was collected in terms of gender, residency program type, and number of years currently certified, though the primary interest for this study is the initial certification year for each examinee. In relation to the diversity of the test-taking population, there are a variety of assumptions made about the wide ranges of ages, ethnic groups, and other identifying demographic criteria for physicians as a population. It is assumed that every physician meets eligibility requirements and maintains a comparable level of medical education and training. Under this assumption the individual differences are not identified as important variables in this study.

Measure

During the real time exam administration, both certification and recertification candidates received identical examination forms, with two equivalent forms of the examination given at each session. The exam is generally administered as a fixed-form, computer-based test with the order of the items within the test form being randomized. The morning session examination content consisted of 120 general questions and two candidate-selected, topic-specific modules of 45 questions each. The afternoon examination consisted of an additional 160 general questions, of which 20 were non-scored field test questions, which were not used for analysis in this study. For the two combined forms there were an initial total of 520 possible questions. Some items were removed for psychometric, content, or editorial reasons, leaving a remaining total item pool of 426 core items, which will be the sample size for this analysis. Due to the content

specific nature of the topic modules and the high degree of individual variation, the modules will be excluded from this study as well.

The exam was developed using concrete psychometric principles and standards; the internal consistency reliability for the exam is estimated at 0.94, which is continually consistent with all ABFM primary certification examinations through the last 10-year period. The exam is designed to measure a single construct, "clinical decision-making within the scope of the practice of family medicine" (ABFM, 2009, p.3). The underlying theoretical construct of the exam design implies that clinical decision-making abilities are the foundation of the ability to recall relevant elements from a larger base of relevant medical information

Prior to analyzing the 2008 examination data, the July 2007 data was rescaled using IRT, Rasch's 1PL model. Using a common-item linkage design, both the 2007 form A and the 2007 form B examinations were placed onto the same scale. Common linkage design provides a method for adjusting the results from one testing instrument to be comparable with another similar testing instrument. This is similar to the scaling used to make comparisons between Fahrenheit and Celsius temperature scales. A formula was derived to transform the ability estimates from the logit metric to a scaled score. The point on the scale where the 2007 minimum passing standard lay (0.6369 logits) was calculated. This value was defined as a scaled score of 390, the 2007 minimum passing scaled score. Similarly, a value of 1.1404 logits was defined as a scaled score of 500. These points were selected to reproduce scaled scores extremely precisely within the range of 390 to 500. Slightly more variation will occur outside of that range. The

formula was designed such that truncating rather than rounding is involved. The formula for converting the raw score Θ to a scaled score SS is:

$$SS = 10 * \text{trunc}\ \{[(218.47 * \Theta) + 250.86]/10\}$$

Again using a common-item linking design, the 2008 form A and 2008 form B examinations were equated to the scale established with the 2007 data. Because the same scale was in place across years, the 2007 passing standard could be retained. The 2007 passing standard was established on the basis of a modified Angoff standard setting exercise. Seventy-one ABFM diplomates who took the examination in 2006 provided ratings for all scored items. Subsequently, the ABFM Exam Committee reviewed the outcome of this project and accepted the final Standard Score cutoff of 390 suggested by the results of the study. In 2008, another modified Angoff standard setting exercise was conducted using 91 ABFM diplomates who passed the examination in 2007. The results of this study suggested retaining the standard (390) set in 2007 and the ABFM Exam Committee accepted this recommendation.

Overall performance on the 2009 maintenance of certification examination resulted in an 82 percent passing rate, accordingly 84.3 percent for the initial certifications and 81.1 percent for the recertification's. In 2007, the passing standard score of 390 was adopted through the process of modified Angoff standard setting procedures, and a repeat standard setting analysis was done in 2008, resulting in the decision to keep the standard set in 2007. The examination committee will review the minimum passing standard just prior to the scoring of the 2011 examination (ABFM, 2009).

Procedure

The data in this study are archival; a cross-sectional design will used in this study. Including within participant information is beyond the reach of this study because of confidentiality, data, record keeping, and time constraints. Therefore, all participants took the certification test at the same time but at different points in their individual careers.

Protection of participants was ensured immediately after the exam data reached the ABFM. ABFM prohibits dissemination of any physician-specific identification data and restricts who can access the data. Therefore, there were no data reported in any way that revealed participant identity. In addition, after the items had been calibrated, examinee information was longer needed or used. The data were sorted by items and participants collapsed into respective cohorts so they could be compared. The only information released was item-level data. Physician identification numbers were not visible nor were they contained in the data in any way. The content of the examination is also prohibited information, so items were not released.

Statistical Analysis

In order to test hypothesis 1 with regards to the correlations of the item calibrations across cohorts, Pearson correlations will be calculated making comparisons of the average item calibrations for all items between the initial certification cohort and each of the recertification cohorts. Cohen's (1992) guidelines for interpreting effect sizes were applied to each correlation to estimate the construct stability of the exam, high correlations will allow for the inference of stability. For the purposes of the study, the

standard of .70 as the indicator of high correlation will be set. Even though it is somewhat arbitrary, this value was agreed upon by subject matter experts as being a desirable cut-off

For the evaluation of hypothesis 2, a Rasch based DIF analysis will be conducted through the Winsteps statistical software program. The item calibrations established for first time test takers will be used with the initial certification cohort, and then recalibrated separately through Winsteps procedures for each recertification cohort. This separation of the calibrations is important to ensuring the stability of the difficulty metric of the exam for the initial and recertification cohorts respectively.

The item calibrations of each cohort will be compared using t-tests to determine any significant differences; these differences will identify the item as showing DIF, or functioning differently for one cohort compared toanother. A t-test will be performed for each item comparison between the initial cohort and each subsequent recertification cohort. Scatter plots of the calibrations for each comparison will be created for a visual representation of the item functioning.

5. RESULTS

Results for Hypothesis 1

Hypothesis 1 predicted that there would be a significant positive correlation between the item calibrations derived from the initial certification cohort and recertification cohort and this correlation would meet or exceed a large effect size as defined by Cohen (1992). Results from the correlation comparing the item difficulty calibrations from the initial certification cohort and the first recertification cohort show the correlation was positive and significant, r (426) = .91, p < .001. The correlation of the item calibrations for the initial certification and the second and third certification cohorts were both positive and significant, r (426) = .81, p <.001, r (426) = .70, p<.001, respectively. Moreover, the subject matter experts agreed that a large correlation meets or exceeds r = .70. Therefore, the first hypothesis was supported.

Results for Hypothesis 2

The second hypothesis stated that there would be fewer than 5% of the items that would demonstrate DIF between initial certification and each cohort of recertification test takers. In order to evaluate this hypothesis, a Rasch based DIF analysis was used. The items were first calibrated for the physicians taking the exam for initial certification, and then calibrated again, separately, for those physicians attempting to gain recertification for each cohort. It is important to note that the calibrations were done separately so the difficulty metric is specific to either initial certification or recertification examinees. The

item calibrations were then compared using *t*-tests. If the *t*-test resulted in a significant difference, the item was identified as showing evidence of DIF.

There were 426 items on the core portion of the certification examination. In order to analyze each of these items for DIF, 426 comparisons were made in the form of *t*-tests. Thus, the preliminary analysis that was carried out was essentially a repeated measures *t*-test with 426 comparisons. Consequently, in order to hold the experimentwise alpha constant (α=.05), it was necessary to adjust for a large number of comparisons. The Scheffé method (1953) was applied to the data to maintain alpha at the .05 level for all comparisons, simple or complex, planned or posthoc, and not just pairwise comparisons as other multiple comparison procedures do (e.g. Bonferroni; Lomax, 1992). The Scheffé method was chosen over the other post hoc adjustments due to the more conservative nature of the method. In Figures 2-4, the scatter plots of the comparisons of the item calibrations are shown with both the initial confidence intervals and the Scheffé to highlight this difference. After applying the Scheffé method, 0 of the 426 items showed evidence of DIF in the comparison between initial certifications and the first recertification cohort. The comparison between initial certification and the second recertification cohort, 5 of the 426 showed evidence of DIF, which equates to about 1.17% of the items. For the final comparison between initial certifications and the third recertification cohort, 16 of the 426 showed evidence of DIF, or 3.75% of the items. Therefore, Hypothesis 2 was supported because well under 5% of the items showed evidence of DIF between the initial certification cohort and the recertification cohort.

6. DISCUSSION

Summary of Results

This study focused on whether the same construct existed for physicians at different points in their careers taking the same certification test. This was done by comparing the difficulty measure calibrations of each item for the initial certification and recertification cohorts using the Rasch model-based DIF. As predicted, the item difficulty measures for each set of cohorts were highly correlated. This is important because a strong positive correlation between calibrations suggests that the test is functioning very similarly across different cohorts. The significant correlation indicates that the test as a whole shows construct stability and the items generally reflect the same pattern of difficulty for both subgroups. As previously stated, a test cannot imply construct validity if it does not measure what it is designed to and have consistent stability over time. This demonstration of construct stability partially supports the validity of the exam.

For the study analysis, the item functioning between the item calibrations of the cohorts gives a big picture analysis of the pattern of difficulty measures. The lack of DIF present in the comparisons further supports the stability of the exam construct. As can be seen in Figures 2-4, the calibrations of the initial certification cohort were plotted on the x-axis, and the calibrations of the recertification cohort were plotted on the y-axis. It is evident by the effect size reported in the results, and illustrated by the scatter plots in Figures 2-4, that there was a strong positive relationship between the difficulty measures of the items between cohorts.

Implications of Results

A strong positive correlation between calibrations of any set of cohorts is essential for certification tests as well as other professional tests. If the correlation was not significant or the direction of the correlation was negative, there would be reason to believe something was inconsistent with the examination or the examinees, again questioning the validity of the exam itself. A negative correlation in this case would indicate that items that were easier for the initial certification cohort were more difficult for the recertification cohort and vice versa. If this were to happen in this study, there would be evidence that a disconnect had taken place between what the subject matter experts who developed the items believed was necessary to know in order to practice family medicine, and what physicians who have recently completed their residency programs have learned. If there was a non-significant or negative correlation between the initial certification and any of the recertification cohorts, it could be that something was missing with the training that the physicians in the initial certification cohort received or that the subject matter experts were not consistent with what training programs are currently teaching.

This disconnect can happen with any certification test as a form of training evaluation and maintenance of the training knowledge and skills. For example, employees in a company may be required go through training and be evaluated soon after training. Then, after a few years these same employees may be evaluated again to ensure maintenance of the information from the training. In such situations, it is imperative that the construct is stable and that the difficulty calibrations of the items show a strong

positive correlation between those that recently completed the training and those that had the training many years prior. This issue should be of concern for many certification tests and other forms of evaluations in order to assess whether the same construct is being measured between members of subgroups at different points in their careers.

The correlation between the cohorts in this study provides insight on the general pattern of item difficulty levels across all items. As this is beneficial information, a closer look at the difference between cohorts at the item level is essential. The Rasch model-based DIF analysis assesses each item individually and how the item does or does not function differently for each career cohort.

The support of the second prediction of this study adds to the understanding of the item difficulty measures between cohorts. The intention of a certification test is to measure a specific construct in order to demonstrate sufficient knowledge on the topic of interest. In this study, the construct should be equivalent for the initial certification and each of the recertification cohorts due to the assumption that physicians need to have and maintain a certain level of knowledge to be certified. Still, it is of no surprise that some items are easier for the initial certification cohort and others are easier for the recertification cohort due to the differences in experience and training between subgroups. When items do show evidence of DIF, as the 21 total of the 426 did in the current study, the questions then are what are the next steps for the analysis of these items at a deeper level?

The seemingly easy answer on the surface to this question would be to eliminate the items in question from the examination completely. If these 21 items are functioning differently between cohorts and the intent of the exam is to measure the same construct,

the items could be considered useless. At a deeper level, this is not a practical answer because eliminating any items that differentiate in difficulty could affect the reliability of the exam. More importantly, the information yielded from an analysis of these items could be useful; the examination of the substantive content of these items may result in an explanation of why these items show evidence of DIF. It is possible that there is something in the item that is typically learned in medical school, so it may be more accessible from memory for those students who recently completed residency. The converse is also possible that the item that is learned through years of practice, and the veteran physicians are more likely to be able to answer correctly. These possible explanations are important to consider when examining the item content in order to understand the difference in functioning between cohorts.

Study Limitations

The greatest limitation of this study is limited sample size, with the use of only one test run of data for comparisons, it is difficult to generalize these results across all testing cohorts. It would be extremely valuable to have the ability to examine the certification exam item functioning across test runs to further investigate the overall stability of this specific certification exam. Valuable comparisons could have also been made using a similar study design with the larger sample of test takers for both exam administrations for the entire year. These limitations open doors for future research and study possibilities to further the field of testing psychometrics.

Future Directions

The results and implications of this study are valuable for many high-stakes tests, certification examinations, performance evaluations, and other assessments that are intended to be equivalent and measure the same construct over time. Certification criteria will always be changing and the knowledge an examinee needs to know will be diverse, but the construct itself should remain stable whether a first time test taker is being evaluated or someone with many years of experience. Future directions for this study may include continuing to examine the functioning of test items with further cohort performance, as well as psychometric evaluation of the specific items showing evidence of DIF. Furthermore, if the examination does contain items that show evidence of DIF between cohorts with varying experience, the in-depth evaluation of the items can give great insight to strengths and weaknesses of subgroups.

This study gives new insight into construct stability across subgroups and training evaluation using physicians who took the ABFM certification/recertification examination. A strong positive correlation was found indicating that the test as a whole functions similarly for all sets of cohorts. A Rasch model-based DIF analysis was conducted to evaluate differences between the initial certification and recertification cohorts. This analysis revealed that 16 of 421 items showed evidence of DIF. Although relatively small, the DIF analysis indicates that further exploration of the substantive content is necessary and could lead to important information about the cohorts and training. Evaluating the differences in item functioning between subpopulations is essential, not only at the ABFM and other medical certification boards, but in many

professional organizations as well. Finding differences and similarities provides insight to the construct stability of the examination and gives feedback for training programs.

List of References

American Board of Family Medicine. (2009). *Certification candidate information booklet.* Lexington, KY.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.). *Educational measurement* (4th ed.) (pp. 221-256). Westport, CT: American Council on Education and Praeger Publishers.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin, 112*, 155-159.

Cole, N.S. (1993). History and Development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings.* Chicago: Rand McNally.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (3rd ed., pp. 201-219). Washington, DC: American Council on Education and National Council on Measurement in Education.

Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Edelen, M.O., Thissen, D., Teresi, J.A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the

likelihood-based model comparison approach: application to the Mini-Mental Status Examination. *Medical Care, 44, 134-142*.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hollman, T. (1973). *Differential validity: A problem with tests or criteria*. Paper presented at Midwest Psychological Association. Chicago, IL. May, 1973.

Institute of Objective Measurement, Inc. (2000). Definition of objective measurement. Retrieved November 10, 2010, from the Institute of Objective Measurement, Inc. Website: http://www.rasch.org/define.htm.

Linacre, J. M. (2008). *A user's guide to Winsteps Ministep Rasch-model computer programs.* Chicago, IL.

Lomax, R. G. (1992). *Statistical concepts: A second course for education and the behavioral sciences.* White Plains, NY: Longman Publishing Group.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955-966.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, 6, 1-3.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. Chapters V-VII, X.

Roever, C. (2005). *"That's not fair!" Fairness, bias, and differential item functioning in language testing.* Retrieved October 22, 2010, from the University of Hawai'i System Website: http://www2.hawaii.edu/~roever/brownbag.pdf

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40,* 87-110.

Stenner, A.J., Smith, M. & Burdick, D.S. (1983). Towards a theory of construct definition. *Journal of Educational Measurement*, *20*, 305-312.

Williams, V. (1997). The "unbiased" anchor: Bridging the gap between DIF and item bias. *Applied Measurement in Education, 10,* 253-267. doi:10.1207/s15324818ame1003_4

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97-116.

Wright, B. D.& Panchepakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement*, *29*, 23-37.

Wright, B. D., Mead, R., & Draba, R. (1976). Detecting and correcting test item bias with a logistic response model. *MESA Psychometric Laboratory, 22.* Website: http://www.rasch.org/rmt/rmt34g.htm.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-like (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research.

APPENDIX A:

Figures

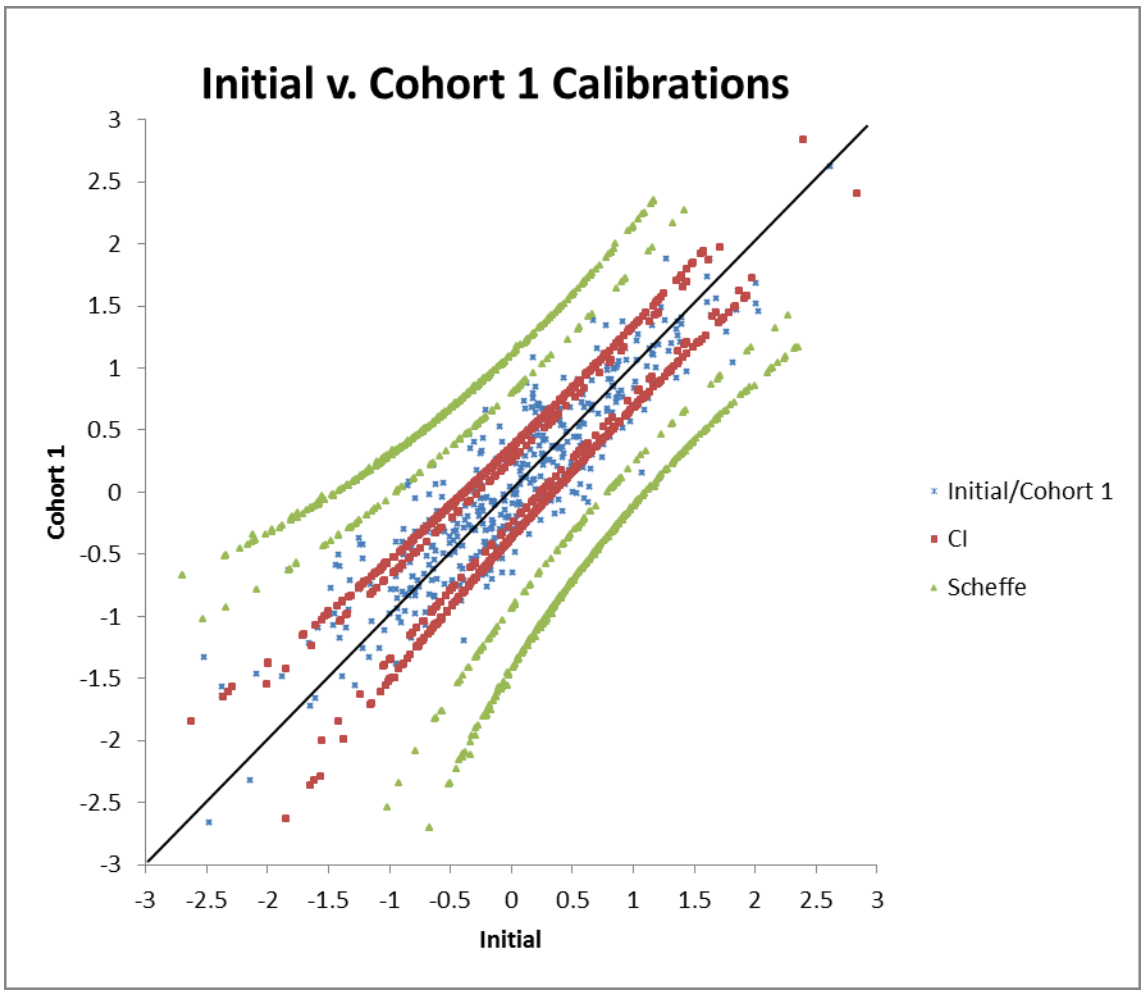


Figure 2. Item Calibrations for Initial Certification Vs. Recertification Cohort

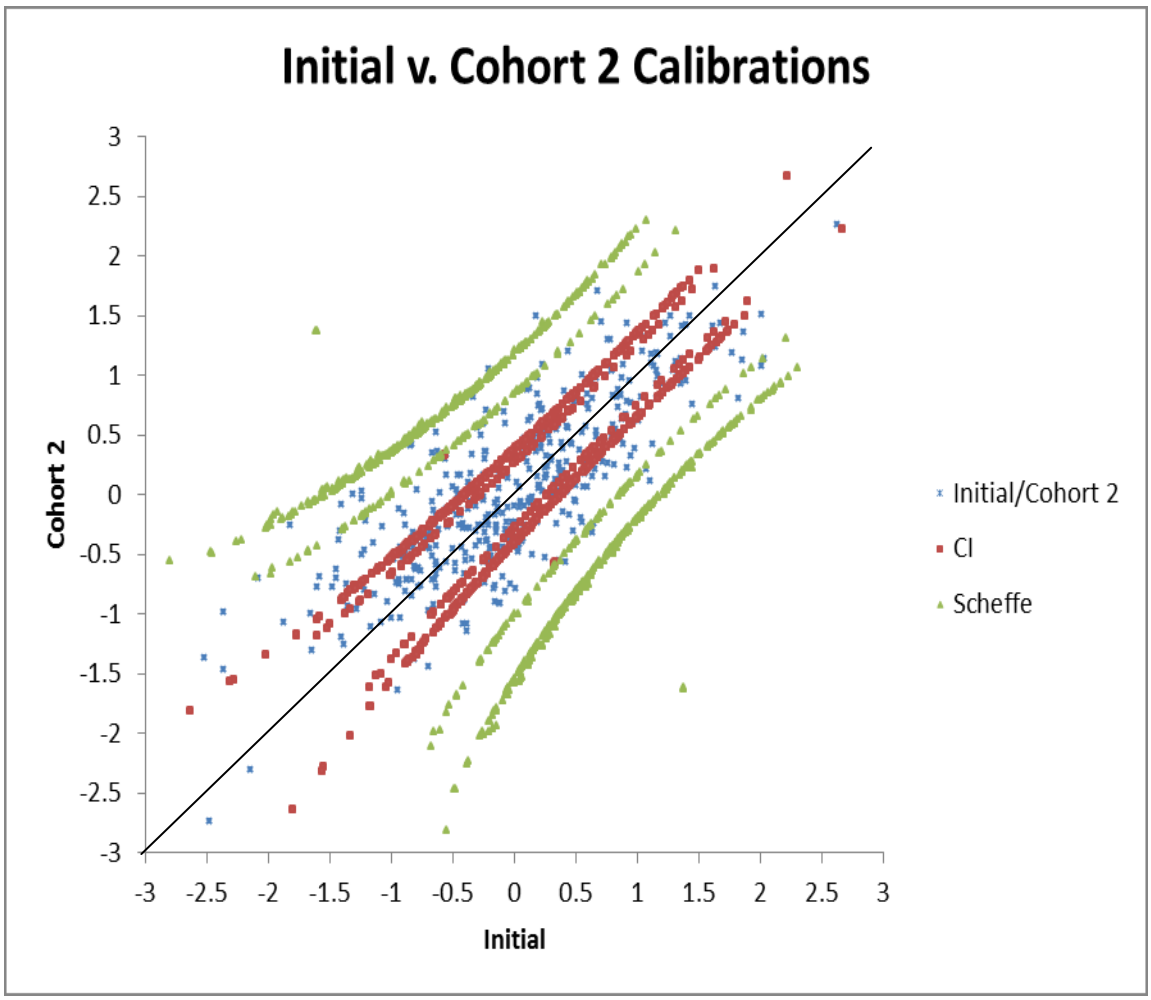


Figure 3. Item Calibrations for Initial Certification Vs. Recertification Cohort 2

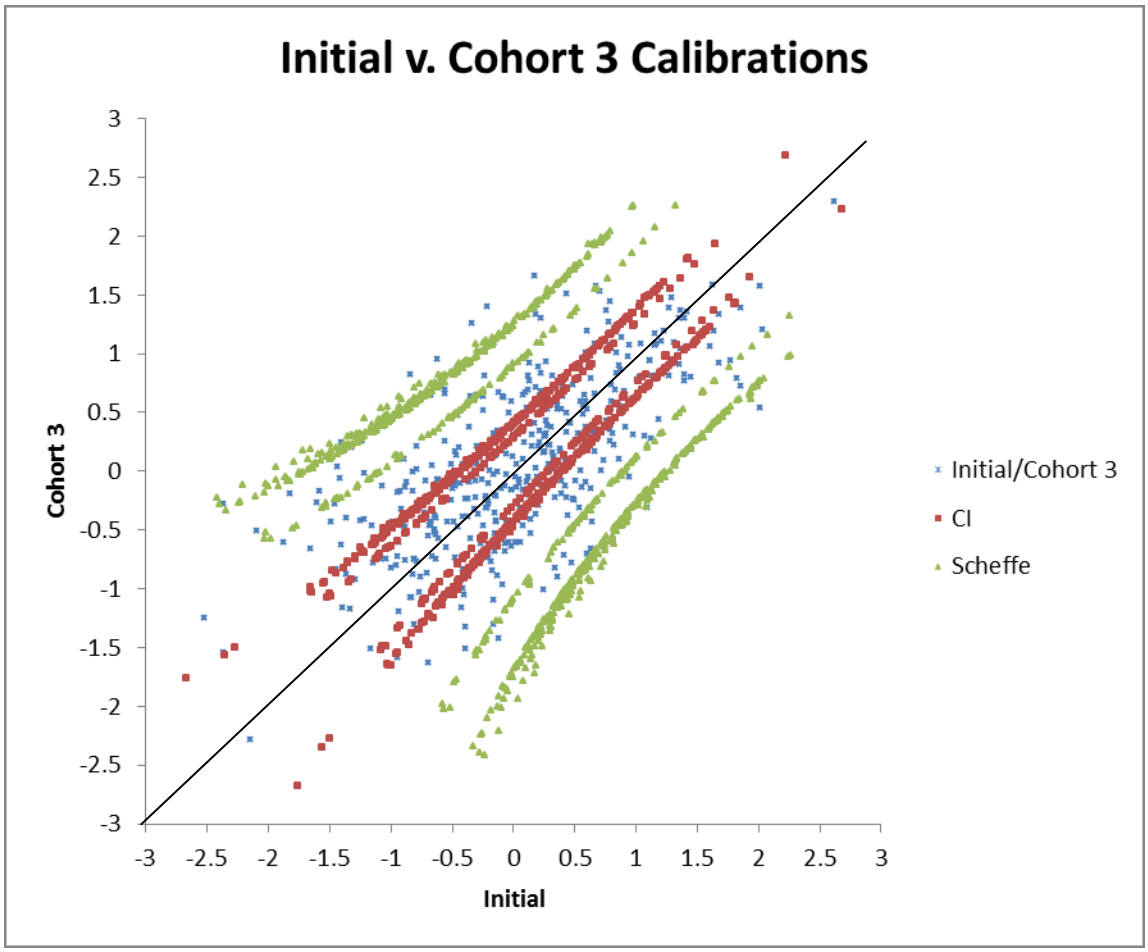


Figure 4. Item Calibrations for Initial Certification Vs. Recertification Cohort 3