

SUNNASSEE, DEVDASS. Ph.D. Conditions Affecting the Accuracy of Classical Equating Methods for Small Samples Under the NEAT Design: A Simulation Study (2011)

Directed by Dr. Richard M. Luecht. 215 pp.

Small sample equating remains a largely unexplored area of research. This study attempts to fill in some of the research gaps via a large-scale, IRT-based simulation study that evaluates the performance of seven small-sample equating methods under various test characteristic and sampling conditions. The equating methods considered are typically applied to non-equivalent [group] anchor test (NEAT) designs using observed scores, where common items are used to link test two or more test forms; that is: (1) the identity method (IDEN); (2) the circle-arc method (CARC); (3) the chained linear method (CLIN); (4) the smoothed chained equipercentile method (SCEE); (5) the smoothed frequency estimation method (SFRE); (6) the Tucker method (TLIN); and (7) the Levine-observed score method (LLIN).

The simulation study design includes 60 test characteristic conditions, including various test lengths and levels of test difficulty and measurement precision, and 20 different sampling conditions related to sample size and the magnitude of ability differences between the samples under a non-equivalent anchor test (NEAT) equating design. The IRT-based simulations provide a powerful way to evaluate equating errors in an absolute sense, even though IRT-based equating is not considered in this comparative study. The ultimate purpose of this study is to establish a set of guidelines that may help testing practitioners better understand which methods of small-sample equating work best

under particular conditions, as well as when small-sample equating may not be appropriate.

The findings suggest that caution is needed when equating small samples under the NEAT design where any of six conditions occur: (1) the sample size for either the base test form or any alternate form is 50 or smaller; (2) the magnitude of the differences in ability between the groups is larger than .1 standard deviation units; (3) the alternate forms differ in mean item difficulty from the base form by more than a quarter of standard deviation unit; (4) the average item discrimination of any alternate test forms is considerably lower than that of the base form; (5) the test forms being equated have too few items (30 or less); and (6) the base form average item discrimination is relatively low. With the exception of these rather extreme conditions, the simulation results suggest that small-sample equating is indeed feasible.

The relative ordering of the seven small-sample equating methods in terms of accuracy (mean bias) is as follows (best to worst): LLIN, CLIN, SCEE, TLIN, SFRE, CARC and IDEN. However all of the methods produce comparable results when the equating samples are similar in average ability. The variability of the equating errors was also used to generally rank-order the seven equating methods, producing the following sequence: SFRE, SCEE, CLIN, TLIN, LLIN, CARC and IDEN. Interestingly, the IDEN and to a lesser extent the CARC methods are consistently most accurate and stable when the equated forms are equal in difficulty (i.e., no equating needed). However, these two methods tend to result in very biased scores for longer tests. Other results were more idiosyncratic in nature and addressed in detail in Chapter IV.

CONDITIONS AFFECTING THE ACCURACY OF CLASSICAL
EQUATING METHODS FOR SMALL SAMPLES UNDER
THE NEAT DESIGN: A SIMULATION STUDY

by

Devdass Sunnassee

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2011

Approved by

Committee Chair

For Mit and the Girls

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of
The Graduate School at The University of North Carolina at Greensboro.

Committee Chair: _____

Committee Members: _____

Date of Acceptance by Committee

Date of Final Oral Examination

ACKNOWLEDGEMENTS

Several people have contributed to the success of my journey in graduate school. First and foremost I consider myself lucky to have had the full and unflagging support of each and every single faculty member in the ERM department. Choosing among them is a very difficult task but I'll start with the members of my dissertation committee.

Dr. Luecht, words cannot describe my deepest appreciation and thanks for all the help and support you have provided me throughout the many years in the program. To this day you remain my source of inspiration and my model as I pursue new avenues in my life. This dissertation would not have been possible without your fierce and unwavering support.

Dr. Ackerman, I'll always remember how you virtually held me by the hands when I started the program. Your faith in me and your optimism and all your support have made me believe in myself. So thank you for all and everything you have done.

Dr. Willse, thank you for all the programming help you provided and for being there when I needed to be brought back to reality in particular during the dissertation phase.

Dr. Chalhoub-Deville, thank you for all the little pep talks I needed to hear. You have helped me see and reach beyond and above myself to realize my full potential. Last but not least, Dr. Morgan, thank you for prompting me into the program when I was still soul-searching and for providing me with the fundamentals in test equating. I would also like to extend my thanks and appreciation to Dr. Henson and Deb Bartz even if they were

not on my committee. They have always been there when I needed them and have provided me with their unconditional support.

In closing, I express my most profound gratitude to my wife, Mit, for her unflinching support and encouragement throughout my graduate education and my dissertation work. Much of the credit for my achievement goes to her and our dearest and lovely girls, Emma, Deesha and our loving Melanie. They have been there for me despite what they had to endure. At this time, I would also like to mention my parents who have instilled in me the virtues of dedication, commitment and patience without which I would probably not have gone and stayed in graduate school in the first place. Finally I would be remiss if I do not acknowledge my family and friends who in their own ways have helped make this journey a reality.

TABLE OF CONTENTS

	Page
LIST OF TABLES	x
LIST OF FIGURES	xii
 CHAPTER	
I. INTRODUCTION	1
Background of the Problem.....	1
Some Specific Challenges of Small-Sampling Equating	5
Statement of the Problem and Research Questions.....	7
Purpose and Significance of the Study	12
Notation and Abbreviations.....	14
 II. REVIEW OF THE LITERATURE	 16
A Definition of Test Equating	16
Properties of Test Equating	16
Equating Designs.....	18
Single Group (SG) Design	19
Random Groups (RG) Design	19
Single Group with Counterbalancing Design.....	20
Common-Item Nonequivalent Groups Design (NEAT)	20
Equating Methods.....	22
Identity Equating.....	23
Mean Equating	23
Linear Equating.....	25
Equipercentile Equating.....	25
Circle-Arc Equating	27
Equating Methods for the NEAT Design	28
PSE (Linear): Tucker and Levine Observed Score.....	28
PSE (Non-Linear): Frequency Estimation Equipercentile.....	31
Chained Equating Type (CE).....	32
Chained Linear (CLIN)	32
Chained Equipercentile (CEE)	33
Levine Equating Type.....	34
Equating Error and Accuracy	34
Review of Past Equating Studies on Small Samples.....	36
Data Collection	38
Equating Methods	38

Livingston (1993) Study	39
Hanson et al. (1994) Study.....	40
Parshall et al. (1995) Study.....	41
Skaggs (2005) Study.....	42
Most Recent Studies	43
Synthesis of Past and Recent Studies	50
Conclusion	55
III. METHODOLOGY	56
Software Resources Employed.....	57
Smoothing Techniques	59
Data Generation and Study Factors.....	61
Sample Size (<i>N</i>)	63
Test Length (<i>n</i>) and Anchor Items	64
Controlling Measurement Precision	65
Base Form Characteristics	66
Unique and Anchor Items	68
Magnitude of Group Separation (STD)	68
Summary of Condition for the Generation of Alternate Forms	69
Data Generation Procedures	70
Equating Steps.....	72
Evaluation of Equating Accuracy and Stability	73
Preliminary Analysis of Base Forms.....	75
IV. RESULTS	79
Effect of Sample Size on Average BIAS	81
Question 1: Accuracy of Equating Methods and av.BIAS	81
Question 2: Interchangeability of Equating Results and av.BIAS	85
Effect of Sample Size on Average RMSD	87
Question 1: Accuracy of Equating Methods and av.RMSD	87
Question 2: Interchangeability of Equating Results and av.RMSD	91
Effect of Magnitude of Group Separation on Average BIAS	95
Question 1: Accuracy of Equating Methods and av.BIAS	95
Question 2: Interchangeability of Equating Results and av.BIAS	99
Effect of Magnitude of Group Separation on Average RMSD	102
Question 1: Accuracy of Equating Methods and av.RMSD.....	102
Question 2: Interchangeability of Equating Results and av.RMSD	106
Effect of Test Difficulty Differences (SMD) on Average BIAS.....	109

Question 1: Accuracy of Equating Methods and av.BIAS	109
Question 2: Interchangeability of Equating Results and av.BIAS	114
Effect of Test Difficulty Differences on Average RMSD.....	116
Question 1: Accuracy of Equating Methods and av.RMSD.....	116
Question 2: Interchangeability of Equating Results and av.RMSD	121
Effect of Discrimination Ratio on Average BIAS.....	123
Question 1: Accuracy of Equating Methods and av.BIAS	123
Question 2: Interchangeability of Equating Results and av.BIAS	127
Effect of Discrimination Ratio on Average RMSD	129
Question 1: Accuracy of Equating Methods and av.RMSD.....	129
Question 2: Interchangeability of Equating Results and av.RMSD	134
Effect of Test Length on Average BIAS	137
Effect of Test Length on Average RMSD.....	140
Final Note.....	143
 V. CONCLUSIONS AND DISCUSSION	 144
Accuracy Summary	144
Accuracy and Sample Size.....	144
Accuracy and Magnitude of Group Separation (STD)	146
Accuracy and Test Difficulty (SMD)	148
Accuracy, a-ratio and Base Form Discriminations	149
Accuracy and Test Length	151
Accuracy and Equating Methods.....	151
Interchangeability Summary	152
Interchangeability and BIAS	152
Conclusion.....	153
Interchangeability and Total Equating Error (RMSD)	154
Study 1: A Special Case	154
Sampling Characteristics.....	156
Test Characteristics	157
Test Difficulty Differences (SMD).....	157
a-ratio and Base Form Discrimination.....	159
Test Length	160
Conclusion.....	161
Recapitulation and Recommendations	161
Equating Conditions	161
Equating Methods	162
Base Form Conditions.....	163
Final Comments and Partial Explanation of Some of the Results	164

Limitations and Directions for Future Research	169
REFERENCES	173
APPENDIX A: DESCRIPTIVE STATISTICS FOR AVERAGE BIAS	182
APPENDIX B: DESCRIPTIVE STATISTICS FOR AVERAGE RMSD	199

LIST OF TABLES

	Page
Table 2.1: Summary of Small-Sample Equating Research and Methodological Studies	37
Table 3.1: Conditions Included in this Study.....	63
Table 3.2: Summary Statistics for Base Form Observed Total, True Total Scores and Anchor Scores in the Population	76
Table 3.3: Summary Statistics of Ability Distributions on Base Form Populations.....	76
Table 3.4: Anchor/Total Correlation and Reliability of Alternate Forms Summarized y Average Item Discrimination and Test Length for each Study	77
Table 4.1: Effect of Sample Size on Average BIAS and SE of Equating Methods for Studies 1-4.....	82
Table 4.2: Effect of Sample Size on Average RMSD and SE of Equating Methods for Studies 1-4.....	88
Table 4.3: Effect of STD on Average BIAS and SE of Equating Methods for Studies 1-4	96
Table 4.4: Effect of STD on Average RMSD and SE of Equating Methods for Studies 1-4	103
Table 4.5: Effect of SMD on Average BIAS and SE of Equating Methods for Studies 1-4	110
Table 4.6: Effect of SMD on Average RMSD and SE of Equating Methods for Studies 1-4	118
Table 4.7: Effect of a-ratio on Average BIAS and SE of Equating Methods for Studies 1-4	123
Table 4.8: Effect of a-ratio on Average RMSD and SE of Equating Methods for Studies 1-4	130

Table 4.9: Effect of Test Length on Average BIAS and SE of Equating Methods for Studies 1-4.....	138
Table 4.10: Effect of Test Length on Average RMSD and SE of Equating Methods for Studies 1-4.....	140

LIST OF FIGURES

	Page
Figure 4.1: Distribution of the 95% CI of the Average BIAS as a Function of Sample Size for the Various Methods by Studies Combinations	83
Figure 4.2: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various Sample Size by Studies Combinations	86
Figure 4.3: Distribution of the 95% CI of the Average RMSD as a Function of Sample Size for the Various Methods by Studies Combinations	89
Figure 4.4: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various Sample Size by Studies Combinations	92
Figure 4.5: Distribution of the 95% CI of the Average BIAS as a Function of STD for the Various Methods by Studies Combinations.....	98
Figure 4.6: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various STD by Studies Combinations.....	101
Figure 4.7: Distribution of the 95% CI of the Average RMSD as a Function of STD for the Various Methods by Studies Combinations.....	105
Figure 4.8: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various STD by Studies Combinations.....	108
Figure 4.9: Distribution of the 95% CI of the Average BIAS as a Function of SMD for the Various Methods by Studies Combinations	112
Figure 4.10: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various SMD by Studies Combinations	115
Figure 4.11: Distribution of the 95% CI of the Average RMSD as a Function of SMD for the Various Methods by Studies Combinations	119
Figure 4.12: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various SMD by Studies Combinations	121
Figure 4.13: Distribution of the 95% CI of the Average BIAS as a Function of Discrimination (a-ratio) for the Various Methods by Studies Combinations.....	126

Figure 4.14: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various Form Discrimination (a-ratio) by Studies Combinations.....	128
Figure 4.15: Distribution of the 95% CI of the Average RMSD as a Function of Discrimination (a-ratio) for the Various Methods by Studies Combinations.....	131
Figure 4.16: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various Form Discrimination (a-ratio) by Studies Combinations.....	135
Figure 4.17: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various Test Length by Studies Combinations.....	139
Figure 4.18: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various Test Length by Studies Combinations.....	141

CHAPTER I

INTRODUCTION

Background of the Problem

For test security or other practical considerations such as time zone differences or practice effects, multiple forms of the same test are administered to different test takers. The forms are commonly referred to as alternate forms and are assumed to have been constructed to the same content and statistical specifications. The test takers may be in the same location all taking the test at the same time or different locations or taking the test on different occasions. A very common example that most practitioners can identify with is the administration of the SAT® exam (The College Board). Multiple forms of the SAT are administered during every testing window, significantly reducing the likelihood that examinees at the same test center can copy or otherwise collaborate with one another. Likewise, alternate forms of the SAT are employed over time to reduce the chance that examinees and test-preparation firms might conspire to memorize and share information about intact, operational test forms.

However, despite every effort to construct multiple test forms that are parallel to one another in both content and statistical specifications, practical limitations related to variation in item writers' interpretation of the test content specifications, the available inventory of items in the item bank for test assembly, and sampling considerations that impact the quality of statistics obtained from item pretesting (tryout) usually result in

producing test forms that differ somewhat—at least in terms of test difficulty (Kolen & Brennan, 2004). One direct consequence of administering forms that differ in difficulty to different test takers is that an examinee’s score will be a function of the difficulty of the form he/she was administered. If one test form is much easier than other forms, examinees taking that form will tend to have observed number-correct or percent-correct scores that are higher than would be expected had he/she taken the more difficult forms. In other words, the observed scores on alternate test forms that differ in difficulty are not interchangeable with one another. If forms are not equated, the use of the unadjusted observed test scores can lead to erroneous decisions that can have serious consequences for the individual or an institution or even at public policy level.

According to Kolen and Brennan, “Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably” (2004, pp. 2). Many testing programs use observe-score equating as a means to address this problem. Observe-score equating can be defined as adjusting the observed number-correct scores on a new version of a test to make them indistinguishable from the scores the test takers would have obtained if they had taken the old version (van der Linden, 2006a). The underlying premise then of any successful equating requires that equated scores have the same meaning regardless of which form, when or to whom the test was administered. Stated a slightly different way, after equating, any examinee ought to be indifferent as to the particular form of a test he or she took.

A perusal of both current and past literature in test equating clearly indicates a long history of test equating and to this day it remains a very active area of research.

Despite almost a century of research on equating, the focus has been geared almost entirely toward large scale testing involving large samples and populations. As a result, virtually all of the equating methodologies that have been developed are based on and are most suitable for large samples. Small-sample equating has not received much attention among researchers and scholars and it is only very recently that research in this area has attracted some interest. In fact, only a handful of studies have investigated test equating in the context of small samples of test takers (Livingston & Kim, 2009).

Without a solid body of empirical research, it remains little more than speculation as to whether or not large-sample equating methods are even appropriate for small samples. Should we blindly accept these methods as adequate for small, possibly non-random samples based on the fact that they appear appropriate for large samples under strict conditions of random sampling (i.e., random assignment of test forms to examinees)? Which methods can we rely on for successful equating of tests based on small samples and under which conditions do they work or fail? Does the accuracy of the different equating methods vary with variations in sampling or test characteristics when sample size is small? These are questions that still appear to have no clear answers in the research literature on small sample equating. They are, however, questions addressed in this study.

One reason for a lack of interest in small sample equating may be related to the notion of standard error as a measure of equating accuracy, typical of all large sample studies. Consistent with the sampling distribution of any statistic, large samples tend to lead to small standard errors (SEs) of estimate and, conversely, small samples tend to

produce very large SEs. Practically speaking, large SEs of equating could suggest unacceptable instability of the equating process—even inaccuracies (Parshall, Houghton, & Kromrey, 1995). This could lead to the rather obvious conclusion that small-sample equating is never advisable because we simply cannot trust the stability of the results. However, the notion that small SEs are the sole criterion for successful equating—or that large SEs are a reason not to equate two test forms—is a prevalent misconception. A small standard error of equating does not always imply a successful equating; it merely indicates a stable result (van der Linden, 1997), and then only if the sampling assumptions hold (e.g., simple random sampling/assignment of test forms).

In practice, at least outside the confines of some of the largest testing organizations, the issue of small sample equating is not uncommon. For example, many certification tests involve very specialized populations (e.g., top-notch computer network engineer trouble-shooters, language experts in Norwegian or Sorani) where tests are administered to only a very small number of test takers—often 30 or fewer examinees in a single year. To make matters worse alternate forms of the tests are administered several times over the course of a year. From a statistical perspective, equating such tests are obviously far more challenging than equating tests using large samples but the inherent issues associated with small samples should not become an impediment to the need for equating test forms in specialized programs in which the number of examinees is small (Parshall et al., 1995). Indeed, this need for applying best-equating practices, regardless of sample size, is strongly supported and outlined in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999 in the context of requiring

practitioners to come with better ways to ensure that scores from alternate forms can be used interchangeably.

Some Specific Challenges of Small-Sampling Equating

Equating small samples presents a very different set of challenges than for large samples. First, because of the small number of examinees taking the test there may not be enough observed scores to cover the full score range on the scale of the test, in particular if the number of test items is larger than the number of examinees taking the test. This problem of restricted-range in the distribution of scores is also possible when the examinee sample is homogeneous¹. In other words, any sparseness of observed scores along the possible score scale introduces into the equating process a missing data set of circumstances (Kim, von Davier & Haberman, 2006). Second, related to the issue of sparse data, is the issue of minimizing the equating error whenever small samples are equated. In large sample equating this issue is not as pertinent because it is well established based on the law of large numbers and the Central Limit Theorem that the accuracy of the equating error decreases with sample size. “If the sample is large and representative, the equating relationship in the sample may accurately represent the equating relationship in the population. The smaller the sample the more likely it is that the equating function computed for that particular sample differs substantially from that of the population” (Kim, von Davier, & Haberman, 2008, pp.325). With small samples the

¹ Note: there is a subtle interaction between the distribution of scores and the measurement precision of a test. For example, mastery, certification, or licensure tests designed to discriminate well in order to classify only a small percentage of the worst or best performers will tend to have a more restricted range of scores overall, than a test designed to best discriminate examinees near the mean.

equating error tends to get proportionally larger, reducing our trust in the accuracy of the equating results. A third, critically important issue is the absence under small-sample equating of an underlying reliable reference scale. For large-sample testing, there is usually a particular “base form” taken by sufficiently large representative sample from the population of interest (e.g., first-takers completing Form 10B in Spring 2011). The reference scale provides a basis for stable comparisons over time². Under small-sample equating, it remains very difficult to choose any particular test form to form the reference scale because the forms are statistically unstable. A fourth consideration is measurement error across the score scale. As van der Linden (2006a) noted, ignoring conditional measurement errors can lead to bias. With large sample, measurement error is usually inconsequential because the extensive amount of “person” information seems to ensure the stability of the score scale³. In small-sample equating contexts, the presence of large measurement error may propagate or augment the sampling errors (of equating) and lead to a worse situation than not equating. Another way of thinking about the issue is to consider often-cited equating assumption that observed scores are random variables because they are for randomly selected examinees from some population (von Davier, Holland, & Thayer, 2004a). Taking observed scores as the variable of interest ignores the existence of random measurement errors and actually underestimates the total error present in any data set: sampling and measurement errors (van der Linden, 2006b). With

² In item response theory (IRT) contexts, the entire bank of operational items is calibrated to a common reference scale. In classical, true-score equating, the reference scale is typically determined by choosing performance on a test form with desired psychometric characteristics.

³ Conditional errors of measurement/estimate, however, are NOT ignorable, even for large-sample equating.

even much less information under small sample conditions, the task of successfully equating such samples gets more formidable.

A final challenge is the choice of statistics as the basis for equating. Most linear equating methods rely on the first two moments of the observed score distributions and their associated interpretations as a measure of form difficulty and dispersion of observed scores. While those interpretations may be appropriate for large--sample equating methods enacted under spiraled forms assignment, this interpretation may be more difficult to justify when equating small samples. Ultimately, the question that we have to ask is whether we should or should not even attempt to equate tests for small samples, and if so, which equating tools might prove to be the most useful under particular conditions of measurement.

Statement of the Problem and Research Questions

This study builds on and extends beyond previous studies on small sample equating. It also attempts to overcome some shortcomings of previous research insofar as comparisons among the various equating methods. There are two classes of shortcomings. First the problem with most, if not all, past and present studies on small or even large samples is that they address the issue of equating accuracy among different equating methods from a resampling approach and focus strictly on ensuring the stability of equating, as indicated by the standard errors. The dependence on resampling is clearly a practical solution for comparing equating results for individual data sets where “truth” is not known-which is obviously the case with real data. Instead, under resampling, truth

is taken to be the expectation of the observed results. That is, the resampled results are averaged and standard errors computed about the mean. There are two limitations with this approach, however. First, resampling does not resolve inherent sampling bias in the original data set(s). Therefore, the empirical expectation of a statistic under resampling could itself be biased. Second, the resampling and estimation of any variance of statistical estimates can mask subtle-but-important conditions that contribute to instability—beyond the sample size. In short, at best, resampling studies provide relative comparisons, not absolute comparisons.

Third, in studies that use real data to make comparisons of the different methods the “true” equating function to which the various methods should have been compared is not known. A review of the literature reveals that there is a somewhat tautological assumption that the equipercentile method is the “gold-standard” criterion metric against which all other equating methods should be evaluated. This assumption has been applied without consideration of sample size (or even sampling mechanisms) While this assumption may be quite appropriate for large sample equating with random assignment of test forms to examinees, is it appropriate to extend the use of this same criterion to estimate the accuracy in small sample equating? The only rationale for choosing the equipercentile methods seems to be that it makes no assumptions about distribution of the scores on the two forms being equated. As noted, while this may be desirable and reasonable for large, representative samples because they usually mimic the characteristics of entire populations, that argument seems less tenable for small samples. In fact, recently, Wang, Lee, Brennan, & Kolen (2008) and Sinharay and Holland (2010a,

2010b) expressed concerns about the choice of a criterion equating method when comparing equating methods because of the possibility that the criterion may favor some methods at the expense of other methods. Unfortunately, it is not currently clear that there is a strong theoretical or empirical basis for choosing the equipercentile or any other particular equating method as the “best”.

Two distinguishing features of this study attempt to overcome the limitations of previous research. First, no particular method is considered to be “best”. In that regard, all of the small-sample equating methods are treated as equal from the onset, where the empirical results determine advantages and disadvantages of each method. Second, using IRT-simulated data where “truth” is known based on a sampled distribution of abilities, this study provides absolute comparisons between the equating methods in terms of both accuracy and stability. That is, this study uses a known “true-score” for each examinee as the basis for comparison of all of the equating methods under investigation under various conditions of measurement. A similar IRT-based approach was advocated by Wang et al., 2008, however, their study focused on large-sample equating methods. By using IRT-based true scores, resampling is not employed nor needed. And this study generalizes to observed-scores as well as estimated true scores.

A legitimate question that might be raised in this study is why choose to compare classical test equating methods rather than IRT equating methods since the data is generated from an IRT model. The reality is that IRT equating was not of interest for this study. The benefits of IRT are best realized when focusing on a calibrated item bank, rather than for form-to-form equating. Examples of where IRT is more appropriate

include computerized adaptive testing or testing programs that employ many test forms within a particular time period. It is extremely difficult and probably not at all cost-effective to develop stable, IRT-calibrated item banks for testing programs that have very small populations. From a practitioner's stand point, form-to-form equating is still most commonly used for small examinees populations. Given the practical need for form-to-form equating, it is important to then realize that consensus of equating research suggests that IRT equating does not have any particular advantage over classical test equating methods when form-to-form equating is used under a non-equivalent, anchor-test (NEAT) design. In addition, classical test equating is generally less complex than IRT equating and makes fewer assumptions (Petersen, 2007). For example, using IRT equating methods, under the NEAT design requires the assumption that all items on the new form, the old form and the anchor test measure exactly the same proficiency. Classical congeneric equating methods relax this assumption. Ultimately, although IRT simulation techniques were used to generate the data, IRT equating was not viewed as being of interest for this study. It should, however, be a topic for future research—a point revisited in Chapter V.

In reality, the IRT data generation/simulation techniques used in this study have four purposes: (1) to provide a convenient way to generate simulated response data following any desired sampling distribution of proficiency and for any desired sample sizes; (2) to control test form difficulty in a very exact way; (3) to control the location and amount of measurement precision associated with the test forms relative to the distribution of examinee abilities; and (4) to compare examinee-level results in an

absolute sense. The latter point needs to be emphasized. Beyond a large-sample study by Wang et al. (2008), this may be one of the first small-sample equating studies to actually know “truth” and provide a clear set of comparisons among equating methods, while at the same time manipulating the quality and characteristics of the test forms in a systematic way.

The broader goal of this study makes a needed contribution to the small-sample equating literature and to formulate some guidelines for practitioners about small sample equating. To achieve these goals, this study examines the accuracy of the equating transformations for different equating methods under various NEAT designs subject to the constraints of small sample size, and including other factors such as test length, group differences, differences between forms in terms of their average difficulty and discrimination (i.e., location and amount of measurement precision). More specifically the major research question addressed in this study can be stated as follows. How do equating group characteristics (sample size and group separation) and test measurement information characteristics (test length, magnitude of discrimination and test difficulty) interact to impact equating results for different methods under small sample, NEAT designs?

This major question is addressed by answering a number of sub questions stated below.

Sub-question 1: How do variations in equating conditions affect the equating accuracy of various equating methods under small sample size conditions under the NEAT design and how consistent are the results with respect to: various sample sizes;

magnitudes of group separation; magnitudes of average test discriminating power; levels of the average test difficulty; and test length?

Sub-Question 2: When are scores of test takers on different versions of a test interchangeable with one another under different equating methods in general, and how consistent are the results across equating conditions with respect to: various sample sizes; magnitudes of group separation; magnitudes of average test discriminating power; levels of the average test difficulty; and test length?

Sub-question 3: Can we establish a set of rules to guide the choice of the most appropriate conditions and or methods when equating can be considered to be successful or fail to work?

It should be noted that the sub-questions 1 and 2 do not address the same questions. The first one addresses the conditions that affect the accuracy whereas the second question addresses the issue of interchangeability of the equated scores among the equating methods. However, these two issues are not necessarily completely independent of one another.

Purpose and Significance of the Study

This study is significant for several reasons. First, the approach to the estimation of absolute accuracy as a means of comparing various equating methods is novel and a rather obvious improvement over the ubiquitous resampling methodology employed for most if not all previous studies on small-sample equating. In this study, equating accuracy is evaluated based on the recovery of the known “true score” corresponding to a

base form, using another equated score from a different test form. Each simulated examinee therefore has a *residual* between truth and the equated observed score, where different equating methods produce different equated scores. The basis for all subsequent comparisons becomes the residuals—which can be aggregated and evaluated conditionally or unconditionally to indicate stability (variances of equating errors) as well as accuracy (indices of bias). Including true scores in the consideration of accuracy also helps disentangle measurement errors from other types of sampling errors. Ultimately, the results of this research are intended to add to the dearth of research on the issue of small samples equating by providing practitioners with tangible guidance as to the propagated effect of a combination of factors such as test length, sample size, the effect size of form difficulty difference or discrimination parameter and group ability differences on the equating accuracy of the various methods—including guidance as to when various equating methods work best or fail to work under particular conditions. For example, can we predict the methods that work best or fail given a set of condition? Is there a method that performs better than other methods across all conditions or is it just under some conditions and if so what are these conditions? These are just some of the practical questions that this study attempts to answer and finding answers to such questions are important to the extent that they can enlighten the practitioner to understand and make sense of equating results when dealing with real data. The results might help clear up some of the uncertainty surrounding small sample equating and lead to making the correct decisions that will cause no harm to the examinees.

In addition, the study findings might confirm or raise concerns about the findings of past studies and bring to light new information about conditions when some methods work best or fail but its importance cannot be overstated. The hope is that the results of this study will help spur more research to establish some standards or parameters against which other designs like the random groups or common item non equivalent groups designs may be evaluated when conducting small samples equating. At present no such operational standards for small samples exist.

Notation and Abbreviations

a: Discrimination parameter

a-ratio: Average Item Discrimination of the alternate form to the average Item
Discrimination of the Base Form

av.BIAS: average bias

av.RMSD: average root mean squared residuals

b: Test Difficulty

CARC: Circle Arc Equating Method

CLIN: Chained Linear Equating Method

CE: Chained Equating

CI: Confidence Interval

CEE: Chained Equipercentile Equating Method

IDEN or ID: Identity Equating Method

LLIN: Levine Observed Score (or Levine Linear) Equating Method

Form X: new form or alternate form

Form Y: old form, reference or base form

FRE: Frequency Estimation Equipercentile Equating Method

P: alternate or new form population

Q: base or old form population

PSE: Post Stratification Equating

RMSE: Root Mean Square Error

SE: standard Error

SEE: Standard Error of Equating

SEED: Standard Error of Equating Difference

SCEE: Smoothed Chained Equipercentile Equating Method

SFRE: Smoothed Frequency Estimation Equipercentile Equating Method

SMD: Standardized Mean Difference in test difficulty between the new form and the old
form

STD: Standardized Group (Theta Difference) Separation

TLIN: Tucker Equating Method

V: anchor item set or anchor test, common items

v: realization of a score on the anchor test

x: realization of a score on the base (old)form

y: realization of a score on the alternate (new)form

CHAPTER II

REVIEW OF THE LITERATURE

A Definition of Test Equating

Equating is a statistical procedure that is used to adjust for differences in difficulty among forms that are built to be similar in difficulty and content (Kolen & Brennan, 2004) and the ultimate goal of equating is to adjust the observed number-correct scores on a new version of a test to make them indistinguishable from the scores the test takers would have obtained if they had taken an old version. By default equating multiple forms of the same test requires that the test forms are parallel in nature. In general two or more forms of an exam are considered parallel when they have been developed to measure the same constructs, are as similar to one another as possible both in terms of the test specifications and statistical criteria.

Properties of Test Equating

Several authors (Angoff, 1971; Lord, 1980; Petersen, Kolen & Hoover, 1989; Harris and Crouse, 1993; Dorans & Holland, 2000) have proposed a number of properties that are important for successful equating to occur and five of these properties are considered as essential:

- The equal reliability Property: The tests should have the same reliability

- Same Specifications Property: Alternate forms are built to the same content and statistical specifications.
- Equity Properties: It must be a matter of indifference to each test-taker whether Form X or Form Y is administered.
- Group Invariance Property: The equating relationship is the same regardless of the group of examinees used to conduct the equating.
- Symmetry Property: The function used to transform a score on Form X to the Form Y scale is the inverse of the function used to transform a score on Form Y to the Form X scale.

In practice the equal reliability and same specifications property ensures that test forms are parallel or near parallel and requires that they are administered under the same conditions of measurement (Dorans & Holland, 2000). The equity requirement, in theory, plays a critical role in test equating. It is important that the requirement be verified after an equating. If the condition of equity is met after the forms are equated then the forms are said to be strictly parallel based on classical true score theory. However, this requirement is hard to evaluate empirically and its use is essentially theoretical (Lord, 1980, Hanson 1991).

Similarly the group invariance property is important but in practice group invariance cannot be assumed to exist in the strictest sense (Kolen & Brennan, 2004). Research in this area by Lord and Wingersky (1984) and van der Linden (2000) have shown that methods based on observed score properties of equating are not strictly invariant, but others Angoff and Cowell (1986) and Harris and Kolen (1986) have suggested the

contrary. In recent years Dorans and Holland (2000) have developed procedures and statistics for evaluating group invariance and equitability.

To the extent that the objective behind test equating is to render scores on alternate forms interchangeable, the symmetry property cannot be overemphasized. If a score x on Form X is equated to a score y on Form Y, then a score y on Form Y, using the same equating method, will be equated to score x on Form X. The essence of this property is that it rules out the use of regression methods for predicting Y-scores from X-scores as a form of test equating because the regression of Y on X is not identical to the regression of X on Y unless there is a perfect correlation between X and Y. Of all the five properties, the equity property is probably the more difficulty one to verify.

Equating Designs

Data collection is clearly the most important aspect of any equating study (Petersen, 2007). The nature of the data collected and the manner in which it is collected will have an effect on the equating process. Several designs can be used for collecting data for equating but most can be classified under three broad categories: the single group design, the random group design and the Non-Equivalent Group Anchor Test (NEAT) design also known as the common item non-equivalent group design. A brief review of some of the designs as described in Kolen and Brennan (2004) is provided in the next section to describe and highlight the conditions for applications and the limitations of some of the data collection designs. For the purpose of this study the focus will be exclusively on the NEAT design.

Single Group (SG) Design

This data collection design requires the administration of two forms of the same test to a single group of examinees. The immediate advantage for this design is that because the same examinees respond to the items on both forms, any difference in performance on the two test forms can be attributed directly to differences in difficulty between the forms. One major criticism of the single group design is that the performance of the examinees is likely to be affected by the order the forms are administered and practice or fatigue effects due to increased testing time. In theory the use of the SG design assumes that the examinees take both forms at the same time in no specific order thereby minimizing if not eliminating the effects associated with order of form administration, fatigue or practice.

Random Groups (RG) Design

This design is also commonly referred to as the equivalent group design where examinees are randomly assigned one of multiple forms. All examinees who receive the same form make up a distinct group. A strict definition of this design requires that all the forms are administered at the same time under the same testing conditions. In practice a spiraling process is used to randomly assign forms such that examinees seated next to one another receive alternate forms and that the examinees are not seated in any systematic fashion (Kolen & Brennan, 2004). The implication of this design is that the difference between group-level performance on any two forms is taken as a direct indication of the difference of the difficulty between the forms with the added benefit that it minimizes testing time relative to a design where examinees take more than one form (Kolen &

Brennan, 2004). The only statistical assumptions that govern this design are that the groups to which the forms are administered are random and independent and that they originate from the same population of examinees. However, one major disadvantage of this design is that the size of the groups or samples has to be large enough for the assumption of randomness to hold.

Single Group with Counterbalancing Design

Under the single group with counter balancing design two forms of the same test are administered to two groups of examinees. The two groups take the two forms in reverse order. If group A takes Form X first followed by Form Y then group B takes Form Y first followed by Form X. The reason behind this is to control for order and fatigue effects which is ignored when two forms are administered under the single group design without counter balancing. The advantage of such a design is that it requires smaller sample size than the random group design since each examinee serves as his or her own control. In practice certain conditions may exist that can make the use of the single group design with counterbalancing more feasible than the random groups design. Such conditions exist if (a) the administration of two forms to examinees is operationally possible, (b) we do not expect differential order effects, and (c) it is difficult to obtain the participation of a sufficient number of examinees in an equating study that uses the random group design.

Common-Item Nonequivalent Groups Design (NEAT)

This design is also commonly known as the Non-Equivalent common Anchor Test (NEAT) design. Two groups of examinees from different populations (P and Q)

are administered test forms X and Y respectively where both forms have a set of common items, V, which is also known as anchor items or anchor test. If V is an internal anchor, then the total test score on Form X and Form Y (expressed as number correct) include scores on V. If V is an external anchor, then total test score on Form X and Form Y do not include scores on V.

One requirement of this design is that the old test and new test forms should be reliable measures of the same construct built to the same test blueprint. A good anchor should ensure a high correlation of scores between the anchor and the total test items on the old and new forms. In other words, the anchor test should behave like a “miniversion” of the old and new test forms. This implies that it should be similar in difficulty to and reflect the content of the unique items on the old and new test forms (Holland & Dorans, 2006).

Unlike the single group or random groups design, the examinees taking any one version of the forms are considered to be non equivalent because only one form can be administered on a given test date (Kolen & Brennan, 2004). This is based on the presumption that every time a form is administered to a group of examinees, the group is not necessarily equivalent to other groups who have taken alternate versions of the same test because the groups may have taken the alternate forms at different times during the academic year when their readiness for this test may be different. Thus differences between the means of two forms can result from a combination of examinee group differences and test forms differences. The central task in equating using this design is to separate the group differences from the form differences. Its main appeal rests on the fact

that it provides information about both group and form differences that the appropriate equating method can take advantage of in transforming scores of the new form to the old form. In other words, this makes the process of equating more rigorous because in adjusting for differences in difficulty that exists between forms the equating method under this design also takes into account differences in ability that might be present between groups taking the alternate forms. Thus different forms administered at different times can be equated based on the 'link' that exists between the forms as a result of the common items they share.

Equating Methods

Equating is a statistical procedure used to transform scores from one form of a test to scores on the scale of some reference form. Several methods exist that can be used to equate alternate forms and they can be classified into two main categories: linear versus non linear methods or classical test theory versus IRT based methods and still others differentiate between the methods as observed score versus 'true score' methods (Dorans & Holland 2000). Some of the most commonly used observed score methods are: mean equating, linear equating and equipercentile equating. The circle-arc (CARC) is the latest addition to the collection of methods and can be classified under the umbrella of non-linear observed score methods. Several other equating methods: Tucker Linear (TLIN), Levin Observed score (LLIN), Frequency Estimation Equipercentile (FRE), Chained Linear (CLIN) or Chained Equipercentile (CEE), have been developed for application under the NEAT design. However, none of them have proved to be entirely satisfactory

and to this day the merits and limitations of these methods continue to be debated because the underlying assumptions upon which these different methods are based and the associated theoretical conditions are not always met in practice.

Some of these methods are reviewed in the next section but a more detailed description of the Circle Arc method is presented in the next section.

Identity Equating

Identity equating is the simplest of all equating method and is generally considered not a “true” equating method because there is no score transformation from the new form to the old form. The new form score is equated to the same score on the old form. Identity equating is no different than no equating. Specifically the identity equating function can be formalized as in equation (2.1)

$$y = IDEN_Y(x) = x \quad (2.1)$$

where x is the realization of the raw score on the new form X , y is the equated score on the old form Y , corresponding to x . The identity function, $IDEN_Y(x)$, converts the new form score x to the old form score y where y is equal to x . The IDEN method is included in this study because it is very common in small sample equating and is very useful when the equated forms are completely parallel.

Mean Equating

For mean equating, the old form and the new form differ in difficulty by an additive constant amount at all points along the score scale. More specifically, to obtain the equated scores, the equating function transforms the scores of the new form by adding

some number of whole or fractional points to the old form. The constant is obtained by taking the difference between the mean of the distribution of scores on the new form and the old form. Thus, the distributions of the equated scores and the old form scores have the same mean but the standard deviation of the equated scores remain the same as the distribution of the new form scores prior to equating. In other words there is a constant shift in the scores of the new form along the score scale but the shape of the distribution is unchanged even after equating. This relationship is formalized in equation (2.2)

$$y = m_Y(x) = x - \mu(X) + \mu(Y) \quad (2.2)$$

where x is a raw score on the new form X , $\mu(X)$ is the mean score for the new form population, y is a score on the old form and $\mu(Y)$ is the mean of the old form population, $m_Y(x)$ is the y score equivalent of x after transformation by the mean function, m . The formula clearly shows that the new form score and the old form score differs by an amount equal to the difference between the mean scores on the two forms.

A consequence of this method is that it does not differentiate between differences in ability of the examinee populations taking the new form and the old form. It makes no difference whether a test-taker is at the high scoring end of the scale or at the low scoring end. The difference in the scores on the two test forms along the score scale remains equal to the difference in the mean scores on the two tests. This is certainly a very unlikely scenario in educational testing. In reality the difference is likely to be smaller than the mean difference for high scoring students if the new form is easier than the old form and much bigger for low scoring students. However, the mean equating is useful to

the extent that it can be easily conceived and visualized in comparing the efficiency and accuracy of other equating methods and is often the method of last resort when the sample size of examinees is too small for other methods to be valid.

Linear Equating

Unlike the mean equating method which assumes a constant difference between the two forms along the score scale, linear equating allows for the difference between the two test forms to vary along the score scale. Linear equating relies on the assumption that the standard deviation (z-scores) scores on the two forms are equal (Kolen & Brennan, 2004). Scores on the two forms which are the same standard deviation from their respective means are considered to be equivalent. In other words the distribution of scores on the new form is transformed to have the same mean and standard deviation as the old form. This relationship is formalized in equation (3) where x is a raw score on the

$$y = l_Y(x) = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X) \right] \quad (3)$$

new form X , $\mu(X)$ and $\sigma(X)$ are the mean and standard deviation of the scores for the new form population, y is a score on the old form, $\mu(Y)$ and $\sigma(Y)$ are the mean and standard deviation of the scores for the old form population, $l_Y(x)$ is the y score equivalent of x after transformation by the linear function, l .

Equipercntile Equating

The definition of the equipercntile equating is attributed to Angoff (1971) and is defined as: Two scores, one on form X and the other on form Y (where X and Y measure the same function with the same degree of reliability) may be considered equivalent if

their corresponding percentile ranks in any given group are equal (p. 563). In other words, if x is the number correct on the new form X and its corresponding percentile rank is P_x then the equated score on the old form Y is the score y on Form Y such that its corresponding percentile rank P_y is equal to P_x . Equipercetile equating is appropriate when the forms differ above and beyond the first two moments. That is the forms also differ in skewness or kurtosis making the relationship between the two forms non linear. In such instances mean or linear equating methods are not appropriate. Unlike mean and linear equating which are strong score models because only a small number of parameters are estimated from the data (Livingston, 1993), equipercetile equating does not make any assumptions about the data nor is it model based. It simply stretches or compresses the raw score units on one form so that the raw score distribution matches that of the other form (Petersen et al.,1989).

However, equipercetile suffers from several drawbacks. First the observed scores on the new form are discrete and finding a corresponding discrete score on the old form can be problematic. Therefore to find the corresponding score on the old form the distribution of scores need to be made continuous and as such some form of *continuization* (Kolen & Brennan, 2004; von Davier, Holland & Thayer, 2004a) is needed to estimate the old form score. The problem is accentuated with small samples because of the sparseness of the data along the score scale. Furthermore, equipercetile equating of small samples is likely to be very highly susceptible to random sampling error (Cook & Petersen, 1987). There is no guarantee that the requirements of successful equating such as equal reliability or population invariance are always met by

equipercentile equating function because they may be violated if the forms being equated are not strictly parallel or equally reliable and the problem may be even worse with small samples.

Circle-Arc Equating

To address the issue of small sample equating, Livingston and Kim (2008) proposed the Circle-Arc method and their results seem to indicate that the Circle-Arc method might be a good alternative to traditional equating methods, in particular when samples are small in size. The original conceptualization of the Circle Arc method of equating is attributed to Divgi (1987); however, Livingston and Kim refined the method. They proposed two versions of the method: the symmetric circle-arc and simplified circle-arc. A full detailed description of the simplified circle-arc method with an application example can be found in Livingston and Kim (2009) and a formal description of the symmetric version can be found in Livingston and Kim (2010). For the sake of clarity, a short description of the main difference between the two methods quoted from Livingston and Kim (2010) is provided below:

Like mean equating, they estimate the entire equating transformation from a single empirically determined point on the equating curve. The curve is constructed to pass through that point, connecting two end-points that are specified without reference to the data. The empirically determined point is the intersection of the mean scores on the test forms to be equated. The upper end-point is the intersection of the maximum possible scores; the lower end-point is the intersection of the lowest meaningful scores. In symmetric circle-arc equating, the equating curve is simply the arc of the circle that includes the two prespecified end-points and the empirically determined middle point. In simplified circle-arc equating, the equating curve is estimated by decomposing it into a linear component (the line connecting the end-points) and a curvilinear component, modeled by a circle arc. (p.176)

Equating Methods for the NEAT Design

Under the NEAT design the common items, V , are used to adjust for population differences. Each examinee comes from only one population and takes only one form. This implies that the NEAT design involves missing data by design (Sinharay & Holland, 2010a, 2010b) and for the equating methods to work, they require strong statistical assumptions (Kolen & Brennan, 2004). Based on different assumptions about the missing data three distinct types of equating methods have been developed to estimate observed score equating functions under the NEAT design (Holland, Dorans & Petersen, 2006; Kolen, 2007). These are the (a) post-stratification equating (PSE), (b) chain equating (CE), and (c) Levine type. The classification of the Levine observed equally reliable method (LLIN) is ambiguous to the extent that some authors classify it in the PSE category along with the Tucker method and others classify it in the Levine type depending on the underlying assumptions. For the purpose of this study the Levine method (LLIN) is classified under the Levine type but for the purpose of reviewing and comparing their assumptions the Levine and Tucker methods are grouped under the PSE (linear) class of methods which starts in the next section.

PSE (Linear): Tucker and Levine Observed Score

Under post stratification equating (PSE), the following non testable assumptions are made: the regression of the total test scores on the new form X on the anchor test, V in the new form population, P is the same as the regression of the total test scores on form Y on the anchor V in the old form population Q . A requirement of PSE is the re-conceptualization of the term “population” because typically an equating function is

defined for a single population but under the NEAT design the scores come from two distinct populations P and Q. To get around this problem Braun and Holland (1982) proposed that these two populations be combined into a new population which they referred to as the synthetic population. The synthetic population is simply a weighted average of the two populations.

Two of the methods in the PSE class which are traditionally referred to in the literature are the Tucker (TLIN) which was first described by Gulliksen (1950, pp. 299-301) and the Levine Observed Score (LLIN) method which was developed by Levine (1955). Both methods are linear methods and apply only to observed score distributions. They both have been investigated under large and small sample equating conditions. In addition to the assumption of the PSE class described earlier, the Tucker method is governed by a second type of assumption. The second assumption requires that the conditional variances of the observed total test scores on the new form Y (or the old form X) given the anchor scores, V is same for both populations to which the alternate forms were administered.

In contrast, the Levine Observed score Model (also commonly referred to as the Levine equally reliable method) makes three assumptions pertaining to true scores which are assumed to be related to the observed scores according to the classical test theory model. The first assumption requires that there is a perfect correlation between the true scores on the total test and the true scores on the anchor test in the old form population and that the same relationship holds in the new form population. The second assumption relates to the linear regression of the true scores on the total test and the true

score on the anchor test. The regression of the total test true scores onto the anchor test true scores is assumed to be the same linear function for both the old form and new form population. In other words, had the new form population taken the old form the regression of their total test true scores onto the anchor test true scores on that test would have been no different than the regression of the new form total test true score on the anchor test true scores. By the same token, the same assumption applies to the regression of the total test true scores to anchor test true scores of the old form in both populations. To summarize, for any alternate form of a test, the linear regression function between the total test true scores and the anchor test true scores on that test is assumed to be invariant within and across populations. As a consequence of the above assumptions the ratio of the standard deviation of true scores on the total test to the anchor test is the same in both the old form population and the new form population.

These assumptions are necessary to obtain estimates of parameters that are not directly observable from parameters that can be estimated. The parameters that can be estimated directly are the mean and variances of the observed scores of Form X and Form Y in their respective populations P (population1) and Q (population2). The not directly observable parameters are: mean and variance of the scores on Form X in the old form population, P, and mean and variance of the scores on Form Y in the new form population, Q. Once these not directly observable parameters are estimated they are used in conjunction with the directly observable ones to estimate the mean and variance of scores of Form X and Form Y in the synthetic population. To obtain the linear transformation of Form X scores to Form Y scores in the synthetic population is defined

by setting the standardized deviation scores equal for the two forms (for more details see Kolen and Brennan, Chapter 4, 2004).

PSE (Non-Linear): Frequency Estimation Equipercentile

A third method under the umbrella of PSE is the frequency estimation equipercentile equating (FRE) method which was first described by Angoff (1971). As in the case of the Tucker and Levine methods, the distributions of scores in the synthetic population are never observed. The technique is based on the estimation of the frequency distributions of Form X and Form Y in the synthetic population. The distributions are estimated using equations:

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x),$$

$$g_s(y) = w_1 g_1(y) + w_2 g_2(y)$$

where $f(x)$ and $g(y)$ are the population distributions for Form X and Form Y scores, respectively. The subscripts s , 1 , and 2 represent the synthetic population, population 1 (new form population) and, population 2 (old form population), respectively. w_1 and w_2 are the weights for population 1 and population 2 that are used to define the synthetic population. However, $f_2(x)$ and $g_2(y)$ are not directly observable.

The frequency estimation is fundamentally an equipercentile method but it requires conditional score distributions to estimate the frequency distributions of both Form X and Form Y in the synthetic population. More specifically it is based on the assumption that the distributions of total test scores on Form X and Form Y given

each anchor, $V=v$, are population invariant. The conditional distributions are formally described by the equations:

$$f_1(x|v) = f_2(x|v),$$

$$g_1(y|v) = g_2(y|v)$$

If we let the marginal distributions of the common items in the two populations be $h_1(v)$ and $h_2(v)$ respectively, then the distributions in the synthetic population can be determined in terms of observable quantities from the equations:

$$f_s(x) = w_1 f_1(x) + w_2 \sum_v f_1(x|v) h_2(v),$$

(see Kolen & Brennan, 2004 for more details)

$$g_s(y) = w_1 \sum_v g_1(y|v) h_1(v) + w_2 g_2(y)$$

Chained Equating Type (CE)

The chained equating type entails three steps. First obtain the equating transformation, t_1 that equates the total test scores on form X to the anchor test V in the new form population. Second obtain the equating transformation, t_2 that converts the anchor V to the total test scores in the old form population. Third to obtain a score y on the old from that corresponds to a given x score on the new form, apply transformation t_1 to x and use the result to then apply t_2 to it to obtain y .

Chained Linear (CLIN). In the chained linear method, the linear transformation, L_1 of the total test scores of Forms X to the common-item V scores in population1 as well as the linear transformation, L_2 of the common-item V score onto the total test scores of

Form Y are first established based on the equalization of standardized deviation scores in each population between total test score and anchor scores. Let $v=L_1(x)$ and $y=L_2(v)$ be these equating functions. Once these relationships are established the composition of the two transformations, $y = L_2(L_1(x))$, (L_2 followed by L_1), will result in the equated score, y in population 2 corresponding to a score x in population 1.

Chained Equipercentile (CEE). In the chained equipercentile equating method, the transformation is based on percentile ranks of the scores in the two populations rather than the moments of the distributions. More specifically, the total test scores (x) in the Form X population is equated to the anchor scores (v) that have the same percentile rank. Similarly the anchor scores (v) are equated to the total test score (y) in the Form Y population. If we let $\Omega_1(x)$ be the percentile rank function in the new form population, P and Ψ_1^{-1} be the inverse of the percentile rank function for the anchor scores in the same population, then the transformation of the total test score to the anchor test in that population is described formally by the equation:

$$v = \Psi_1^{-1}\Omega_1(x)$$

If we let $\Psi_2(v)$ be the percentile rank function of the anchor scores in the old form population and Ω_2^{-1} be the inverse of the percentile rank function total test score in the same population, then transformation of the anchor test to the total test in that population is described formally by the equation:

$$y = \Omega_2^{-1}\Psi_2(v)$$

The composition of the two equipercentile transformations involved in equating scores x in the new form population (population1) to y in the old form population (population2) is described formally by equation:

$$y = \Omega_2^{-1}\Psi_2(\Psi_1^{-1}\Omega_1(x))$$

Levine Equating Type

The Levine equating type, is based on the assumption related to the true scores between the total test and the anchor in each population. More specifically, the Levine Equating methods assume that the true scores on the new form X and the anchor test V in the new form population are perfectly linearly correlated and that the same applies to the true scores on Form Y and V in the old form population.

To the extent that the Levine-observed score equating method uses estimates of true scores in determining the mean and standard deviations of the new and old forms on a common synthetic population it can also be classified as a Levine type method. In general the major difference between the Levine type methods and the other types is that it relies on the estimation of true scores as defined in classical test theory.

Equating Error and Accuracy

The very nature that equating is a statistical process dictates the quality of the equating results. One way to gauge the quality of any equating procedure is to evaluate its accuracy as a function of equating error. Equating error can be decomposed into two sources of error: random error and systematic error. These two sources of error are

unavoidably present in an equating study. The root cause of random error can be attributed to the fact that samples and not the whole population are used in estimating the equating relationship from the new form to the old form. If the whole population is used then no sampling error would be present. On the other hand, systematic error can be attributed to one or more factors ranging from the idiosyncrasies of the equating method, the degree to which the assumptions of the method are met or violated, the data collection design, alternate forms that are not entirely parallel or even large group differences between the new form group and old form group. Increasing the sample size will reduce the random error but will have no effect on the magnitude of the systematic error components (Kolen & Brennan, 2004). By extension these measures of accuracy are bound to have direct and much stronger implications for the evaluation of the discrepancies that are present in small samples equating.

Two school-of-thoughts currently exists as to what constitutes equating error. The first school seems to share the notion of accuracy in terms of random error based on the equating transformation of the marginal distribution of scores at the sample level and does not seem to be too much concerned with measurement error or systematic error. This concern is expressed in the following quote from van der Linden (1997): “Adjusting (marginal) observed-score distributions to one another does not tell us much about what happens to the distributions at the level of the individual examinees” (pp. 289). This measure of accuracy is usually expressed in the form of the standard error of equating.

The other school takes the opposite view that systematic error expressed in the form of bias and measurement error at the level of the examinee are more important

factors affecting the accuracy of equating methods. Indeed the position of the proponents of the second group is summarized in the quote from van der Linden (2010):

The equating literature has been dominated by an interest in the standard error of equating, but bias is the primary criterion for evaluating the success of an equating. After all, equating is an attempt to remove the bias in the score on the new test form as an estimate of the score on the old form due to scale differences. A focus on the standard error of equating prevents one from noticing any remaining bias in the equated scores, or even possible new bias added to them in the equating process. (p. 21).

Review of Past Equating Studies on Small Samples

A plethora of articles on equating in the literature spans several decades, however, only a handful of studies have been conducted on small samples equating to date. The past five years has seen a little surge in the number of publications related to the topic. Livingston and Kim co-authored the two most 2010 recent publications on this subject. Other recent publications include Kim et al. (2008), Heh (2007), Skaggs (2005). Parshall et al. (1995), Hanson, Zeng, and Colton (1994), Livingston (1993), and Harris (1993) are among the most notable publications in the last decade. A review of these articles indicates that like in the large sample case some methods are more appropriate than others depending on the relationships of the scores between the forms to be equated. At the same time the choice of an equating method is closely linked to the equating design used to collect the data.

Table 2.1 below provides a summary of some of the most recent studies that have been conducted on the topic of small sample equating. The table shows the data collection design, the sample sizes examined, the equating methods, the number of pairs

Table 2.1. Summary of Small-Sample Equating Research and Methodological Studies

Study	Design	Sample sizes	Method(s)	Pairs of forms equated	Criterion
Livingston (1993)	Random groups with common items	25, 50, 100, 200	CEE with and without pre-smoothing	1	Equipercntile
Hanson et al (1994)	Random groups	100, 250, 500, 1000, 3000	Equipercntile with various pre- and post-smoothing methods	5	Equipercntile including model based fitted distributions
Parshall et al. (1995)	NEAT	15, 25, 50, 100	Levine-Angoff (linear)	5	Levine-Angoff (linear)
Skaags (2005)	Random groups	25, 50, 75, 100, 150, 200	Equipercntile with and without pre-smoothing, linear, mean	2	Equipercntile with 6-moment log linear presmoothing
Kim et al. (2008)	NEAT	10, 25, 50, 100, 200	Identity, Chained Linear, Synthetic	2	Chained Linear
Livingston & Kim (2009)	NEAT	25 for new form and 75 for reference form	Identity, Mean, Levine (linear), Tucker, Simplified Circle-Arc, Chained Linear, Chained Equipercntile	2	Chained Equipercntile with presmoothing
Livingston & Kim (2010)	Random groups	50,100, 200, 400	Equipercntile with smoothing, Linear, Mean, Symmetric Circle-Arc, Simplified Circle-Arc	6	Equipercntile
Kim & Livingston (2010)	NEAT	10,25,50, 100	Identity, Chained Linear, Chained Mean, Symmetric Circle-Arc, Simplified Circle-Arc, Chained Equipercntile with smoothing	4	Equipercntile

of forms compared and the equating criterion used in each study. It is clear from these studies that there is no general consensus over what actually constitutes a small sample or what size would constitute the lower and upper limit of a “small” sample. Samples as small as 10 (Kim et al., 2008) up to 3000 (Hanson et al., 1994) have been used in these studies with sizes ranging from 25-200 appearing to be the most common among these studies.

Data Collection

The table also shows that there have been more studies on the accuracy of equating methods that focused attention on the NEAT design than on the random groups design. Two major reasons may be attributed to this choice of data collection design on the part of the researchers. First the NEAT design does not require as large a sample as the random groups design (Skaggs, 2005). The second and probably the more important reason is that to maintain test security in high stakes testing many of the testing programs want to give a new test form at every administration and therefore do not want to use to use an equivalent groups design because it requires re-administration of an old test form for equating purposes (Petersen, 2007).

Equating Methods

Table 2.1 also shows that a variety of different methods (linear and non linear) have been examined in those studies. The linear equating transformations include the Identity, Mean, Chained Mean, Chained Linear, Tucker Linear, Levine Observe, Synthetic and the non-linear methods include Equipercentile, Chained Equipercentile with and without smoothing, two versions of the Circle-Arc: Simplified Circle-Arc and

Symmetric Circle-Arc. The first observation of interest is that all these methods can be classified under the observed score models. Second, the Equipercentile with or without smoothing is the only one method that has been studied across all the studies. However, the Equipercentile method is not necessarily suitable in small sample equating studies because of the high potential for unrepresentative samples. Had it not been for the recent advent of the Circle-Arc, the Equipercentile method would have remained the only non-linear method studied in small sample equating. This is in stark contrast to the range of methods (Frequency Equipercentile Estimation, IRT-based) that have been developed and used in large sample. Third the Equipercentile equating in small samples require that some smoothing procedures be used to ensure stable results, even more so than in large samples because of both the sparseness and discreteness of the data. Livingston (1993), Livingston & Kim (2009, 2010), Skaggs (2005) have applied presmoothing procedures based on the loglinear model to the discrete data prior to conducting the Equipercentile transformation of the scores on the new form to the old form.

Livingston (1993) Study

Livingston (1993) examined the effectiveness of log-linear pre smoothing (no smoothing, 2, 3 and 4 moments) with samples of 25, 50, 100, and 200 examinees per form using the Chained Equipercentile method under the random groups design with common items. He conducted a resampling study and equated two forms of the Advanced Placement History Examination which differed by three-fourths of a standard deviation. He found that the benefits of smoothing were greatest when the sample was small. Specifically he found that presmoothing significantly reduced equating error, more so for

the smallest samples. He found that equating error using presmoothing was about as effective as unsmoothed equating using samples twice as large (Skaggs, 2005). von Davier (2007, p.97) concurs with these findings but she adds that the sample characteristics may affect the number of moments that should be preserved in the observed distribution. The number of moments in the observed distribution that should be preserved in the smoothed distribution depends on the sample characteristics and the major caveat with log linear smoothing is that it may introduce a large enough sampling bias in small samples which can offset any gain made in the reduction of the standard error of equating (Livingston, 1993).

Hanson et al. (1994) Study

Hanson et al. (1994) studied various equating methods, identity, linear, unsmoothed, presmoothed, and postsmoothed Equipercetile equating under the random groups design for five ACT Assessment Tests with samples ranging in size 100 to 3000 observations. They conducted a resampling study and their findings indicated that, identity equating resulted in less equating error than any other linear or equipercetile method with samples of size 100. There was no obvious preference between pre-smoothing and post smoothing method though smoothing significantly improved the equipercetile equating with small samples. Skaggs (2005) reported that in an extension of the study by Hanson et al, Kolen and Brennan (2004) compared identity equating to the other methods. They found that when the forms differed substantially in difficulty (.6 SD difference) the total equating error was greater than when the forms were similar in difficulty (.1 SD difference). Even when the forms were similar in difficulty identity

equating produced less error than any method for samples of size 100. However, when the forms differed, some of the methods had less equating error than identity equating at some points on the raw score scale.

Parshall et al. (1995) Study

In probably the first small sample equating study under the NEAT design, Parshall et al. (1995) examined standard errors and statistical bias as indicators of accuracy in samples of size, 15, 25, 50 and 100 using the Levine-Angoff (Angoff's Model IV) linear equating method. They conducted a resampling study by drawing small random samples from state teacher certification tests in five subject areas, calculating the equating transformation derived from each sample, and evaluating these functions for statistical bias and sampling error. The difficulty differences between pairs of forms ranged in effect size from zero to 0.4 of a standard deviation. The groups (pseudo populations) from which samples were drawn were practically equivalent in ability. They found that sample size affects the standard error of equating with smaller sample size leading to larger standard error. Furthermore as raw scores deviated further from the mean the standard error increases monotonically, leading them to recommend caution when equating passing scores that are distant from the mean. Another important finding of their study is that the tests that had the highest percentage of overlap and highest correlation between anchor and total tests had the least amount of equating error.

Skaggs (2005) Study

Skaggs (2005) study on small sample equating had two objectives. The first focused on the accuracy of equating small samples under the random groups design with no anchor and the second related to the impact of the accuracy at the passing score. He compared mean equating, linear equating, unsmoothed equipercentile equating, and equipercentile equating using two through six moments of log-linear presmoothing with samples of 25, 50, 75, 100, 150, and 200. He used data from the Social Studies Test of the Tests for General Educational Development (GED) and two alternate forms of the test. The two forms consisted of 50 multiple choice items and were relatively close in difficulty differing only by about 1/10 of a standard deviation. They both had reliabilities 0.9 or greater. Like in previous such studies he used a resampling procedure where samples for different sizes were drawn without replacement for each test from from the populations of the two forms. To evaluate the equating results he computed the equating bias, the standard error of equating and the root mean squared deviation at each raw score point of the score scale. He used a large sample criterion, in this case equipercentile equating of the populations of the two forms, to compute these statistics. Like (Parshall et. al, 1995) he found that the standard errors of equating decreased with larger sample size and that increasing sample size reduced bias but the improvement varied across methods. He recommended not equating for samples of size 25 or smaller because of equating under such conditions are likely to cause more harm to the examinees. For samples 50 and above some form of equating is preferable to no equating and that presmoothing using the log-linear models that fit the first two or three moments of the

observed distribution tends to lead to smaller standard error than did unsmoothed equating. Higher levels of presmoothing beyond three moments did not improve the results.

Most Recent Studies

The Kim et.al (2008) study is probably the first study that introduced a method specifically developed to tackle the issue of equating in small samples. The method is a compromise between the identity function (no equating) and an estimated equating function. They called it the synthetic linking function (syn) which is defined as the weighted average of an estimated equating function and the Identity function ($IDEN(x)=x$).

$$syn_Y(x)=we_Y(x) +(1-w)IDEN(x),$$

where w is a weight between 0 and 1, $e_Y(x)$ is the estimated equating function.

Specifically they compared the identity equating, chained linear and synthetic functions with a weight w of 0.5 for samples of size, 10,25,50,50,100 and 200 under the NEAT design. The data for this study came from two national assessments: one with a highly reliable external anchor (0.84) and the other with a moderately reliable internal anchor (0.67). The effect size of the difference between the means on the anchor test of the two alternate forms in the first assessment was (0.32), which in the authors' opinion is fairly large. On the second assessment the effect size was only (0.05), indicating a negligible difference between the two populations taking the alternate forms for this assessment. The correlation between the anchor test and the total test was pretty high

hovering around 0.87 or 0.88 in all four populations. This study like the previous ones used a resampling procedure to evaluate the equating accuracy except that they chose the chained linear equating function as the criterion function. They argued that based on the studies by Harris (1993) and Kolen and Brennan (2004), the samples sizes are too small to ensure the adequacy of equipercentile equating as the criterion equating function. Like in the previous studies they computed bias, SEE and RMSE at each score point along the score scale as measures of accuracy. In addition they also computed the standardized root mean squared error (SRMSE) which returns a single statistic which is the RMSE averaged over the entire score range and weighted by the standard deviation of the scores of the old form population.

Their findings indicated that when sample sizes are small (fewer than 50), the synthetic function outperformed the chained equating function because the identity function was part of the synthetic function. They suggested that the synthetic function method may be an alternative when sample sizes are small and groups differ in ability. Even for samples as large as 200, the synthetic function or the identity function outperformed the chained linear equating method regardless of anchor quality. Although its bias was relatively smaller the chained linear method showed the greatest amount of equating error. They concluded that for well designed and almost parallel forms, the identity function is likely to do less harm than conventional equating and may not be appropriate when these conditions do not hold. They support Skagg's recommendation to use the identity function when the forms differ by one-tenth of a standard deviation or less. In their final remarks they cautioned that the method should not be used blindly

because the symmetry property required in equating is a function of several factors such as the choice of estimated equating function, the choice of weights, the condition that the ratio of the total score variance to anchor score variance in the old form population is equal the ratio in the new form population.

Another method, the Circle-Arc, specifically designed to equate small samples was introduced by Livingston and Kim in their 2009 study. Like earlier studies on small samples this study basically followed the same procedures except for the inclusion of the new equating method in the list of other methods to which it was compared. They conducted the same traditional resampling procedure where a pair of test forms is equated in large groups of test takers and then in repeated small samples of size from the large groups. They used samples of size 25 for the new form and based on their experience in equating small-volume tests, they chose samples of size 75 for the old form. They examined several methods including the Chained Equipercentile, Levine (observed), Chained Linear, Tucker, Mean, Circle-Arc and Identity under the NEAT design. The criterion equating was the chained equipercentile transformation with smoothing of the new and old form populations. They applied the log-linear model that preserved the first five univariate moments of each marginal distribution (i.e., of the total score and the anchor score) to presmooth the joint distribution of the total scores and anchor scores in each population. For the chained equipercentile equatings of the small sample scores, only the three univariate moments of each marginal distribution were preserved in the presmoothing process. They used data from two alternate forms of a multiple choice test that is widely used for the certification of prospective teachers. The two populations

taking each form were about the same ability (0.03 SD difference on the anchor) but the effect size between the average score on the full test was 0.36 SD indicating substantial difference in the average difficulty between the forms. The correlation between the total and anchor scores was high on both forms, hovering at 0.90 for both of them. For the evaluation of equating accuracy they used bias, standard deviation of equated score (SD) and root mean square difference (RMSD) at each score point along the score scale that ranged from 30 to the maximum score of 107. For the Circle-Arc method the minimum score on the test was set at 30 because they determined that this is the score that would be expected by chance or guessing. For scores below the chance level, they suggest using a linear function (generally the identity) to connect the chance level scores to the point representing the minimum possible scores on the two forms.

Their findings indicated based on the RMSD all the methods had about the same accuracy for scores that were near the median of the score distribution but differences between methods increased in monotone fashion for scores further away from the median. They found that overall the Circle-Arc was the most accurate method and that methods based on very strong assumptions (mean equating and Circle-Arc equating) were more accurate than the other methods that are based on weaker assumptions.

With respect to bias, it was clear that equating bias was a function of both the equating method and the location of the score along the score scale. All the methods except the identity equating exhibited relatively smaller degree of bias in the middle of the score scale. As with the RMSD the bias increased monotonically for scores further away from the middle section for all linear methods. The two non-linear methods (Circle-

Arc and chained equipercentile) exhibited bias in opposite directions outside the middle of the distribution.

In terms of the variability of the equated scores expressed in terms of the standard deviation of equating, all the methods produce about the same variability (0.1 SD units) at or around the mean. For scores that deviated further away from the average in the raw score distribution, the methods based on weak assumptions exhibited increasing variability (as large as 0.5 SD units) associated with the random component of the equating methods. If we ignore the scores at the end of the scale, the variability in the equated scores for the Circle-Arc method was the lowest among all the methods, hovering at 0.1SD units in the middle of the scale and decreasing towards the tails of the distribution. The mean method was the second best with a constant variability of 0.1 SD units along the entire score.

Livingston and Kim (2009) concluded that the Circle-Arc could replace mean equating as the choice for small sample equating and that it has the advantage of being more accurate at the upper and lower tails for the score distribution if pass/fail decisions are made at cutscores that lie at the tails. They also suggested that when forms differ in difficulty and samples are too small for equipercentile equating, Circle-Arc may be an alternative solution.

Kim and Livingston (2010) is a follow up study to their 2009 study on the Circle-Arc method of equating in small samples under the NEAT design. In addition to chained equipercentile with smoothing and chained linear equating methods, they included the chained mean equating method and two versions of the Circle-Arc equating method

namely the simplified Circle-Arc and the symmetric Circle-Arc. Again they used the same resampling procedures that are customary with small samples equating, except that they equated four pairs of test forms that were constructed from four operational forms that had at least 110 items each and were administered to more than 10,000 examinees.

The four pairs of research forms had 71, 70, 69 and 63 items each and shared at least 23 items in common. They were equal in length and parallel in content and the correlations between the total score and the anchor score ranged from 0.85 and 0.91. The forms however, differed in difficulty ranging from 0.17 to 0.30 standard deviations based on the data from the combination of the two populations, each taking one of the pair of forms.

In addition, the four pairs of examinee group ability ranged from -0.30 to 0.30. They performed a total of 32 resampling studies that is a combination of four pairs of forms, two assignments of examinee populations to test forms and four specified samples sizes. The samples sizes used were 10, 25, 50 and 100 for the new form and three times these sizes (30, 75, 150, 300) for the old form samples.

To evaluate the accuracy of the equating methods they conducted a direct equipercentile equating of the new form to old form scores formed by the combination of the two examinee populations from which the equating samples were drawn. The statistic used in this study was the RMSD at each raw score point of the score scale. By conditioning the RMSD values on the percentiles of the new form distribution and expressing them in terms of the standard deviation of the distribution, they combined these values across each of the four sets of eight of resampling studies (eight

combinations of test form pair and group assignment). They averaged over the eight resampling studies at each of nine different percentiles (1%, 5%, 10%, 25%, 50%, 75%, 90%, 99%) to compute the RMSD at these points.

Their findings indicate that overall the chained mean and the two versions of the circle-arc produced the smallest equating error along the score scale. For very small samples of size 25 or fewer on the new form and 75 or fewer on the old form, the two versions of the Circle-Arc method produced the smallest equating error. For samples of size 100 on the new form and 300 on the old form the two methods proved to be the most accurate at the higher end of the score scale. For scores at or below average the chained mean method seemed to be more accurate than all the other methods across all sample sizes except for the case of sample size 100 on the new form. At samples of size 100 on the new form and 300 on the old form the chained mean method was no longer a match for the other methods except for scores at the center of the distribution where they performed equally well. It also performed relatively poorly for scores above the 50th percentile across all sample sizes.

Both the chained linear and equipercentile methods produced large equating errors below the lower and above the upper quartiles with sample size smaller than 25 on the new form and 75 on the old form. As would be expected the equating error produced by these two methods were even larger in the case of sample size 10 on the new form and 30 on the old form. However, the identity equating did not prove to be that much better than the other methods for scores below the 75th percentile even for samples of size 10 on

the new form. The authors attribute this observation to the fact that the forms were constructed to differ in difficulty to ensure that equating would become a necessity.

In conclusion the Kim and Livingston (2010) study confirmed the findings of their previous study where the Circle-Arc method proved to be a good alternative solution to equating in small samples in particular when samples tend to be very small (50 or less and when decisions about pass/ fail designations have to be made at certain specific scores on the score scale. It also showed that the two versions of the circle are no different from one another.

Livingston and Kim (2010) conducted another similar, counter-part study to their small samples equating study under the NEAT design. This time they compared the accuracy of various equating methods under the random groups design. A comparison of the two studies reveal that in many ways, whether in terms of the methodology or the measures of equating accuracy or the use of the Equipercentile method as the criterion equating function the two studies were practically similar. Furthermore their findings for this study were very similar to those of their (2010b) study even if the sample sizes used ranged from 50 to 400 and were larger than those used under the NEAT design.

Synthesis of Past and Recent Studies

A number of common themes are evident from the studies that have been reviewed. First all the studies involved some form of resampling of small samples from a larger population where the number of samples (iterations) ranged from 200 in Skaggs studies to 1000 in Kim et al. (2008) and others. The criterion equating function was

always some form of an observed- score equating function and for the most part the direct equipercentile equating function is the preferred one. However, it does not appear that there are strong theoretical justifications in the literature why it should be preferred to other methods used as the criterion for evaluating equating accuracy other than it does not make any assumptions about the marginal distributions.

Second most of the studies are based on real data and under such circumstances there is a “generally a lack of clear criterion (or true equating function) for evaluating equating bias” (Sinharay & Holland, 2010a, 2010b). For example Kim et al. (2008) used the Chained Linear Equating methods as the criterion arguing that their choice is dictated by the need to “... avoid confounding the differences in accuracy with the differences in shape between large and small samples” (pp.328). Some of these studies include the same or a version of the method used as the criterion among the other methods being investigated. The issue with this approach is clearly spelled out in Wang et al. (2008) where they stated that such a situation may give an undue advantage to the method that is being examined and used as the criterion at the same time.

Third the evaluation of accuracy is generally based on three statistics: bias, SEE and RMSE. Kim and Livingston tend to prefer SD for SEE and RMSD for RMSE. These statistics are generally computed at each score point of the raw score level. So for a 30 item test, these statistics will be computed at each score point starting from 0 to 30.

The bias at a given raw score point is defined as the discrepancy between the equated score for that score point and the corresponding criterion score averaged over the number of iterations in the resampling procedure. The SEE at a given raw score point for a

number of iterations is defined as the standard deviation of the equated score for that score point and the RMSE at that point is simply the square root of the sum of the squared bias and variance. The concern with these definitions is that the equated score is never compared to the old form score or to “truth” which is generally known in simulation studies.

By definition, equating requires that an examinee’s equated score and the corresponding score on the old form should be indistinguishable. The accuracy of an equating method should therefore be judged in terms of the discrepancy between these two scores. However, with real data under the NEAT design the old (new) form score of a given examinee from the new (old) form population is missing because it is never measured. As such the authors of these studies do not have the luxury of knowing the actual “true” discrepancy and therefore resort to a criterion as a substitute for the “true” equating function. The problem with this approach is that the examinees’ observed scores are:

...pooled into a population distribution for each of the two groups in the equating study, whereupon the two distributions are redefined as distributions for a synthetic population, from which a single equating transformation for all examinees is derived. This pooling and synthesizing of observed scores for individual examinees, along with the derivation of a single transformation, may be the most serious source of bias in traditional score equating. van der Linden (2010, p. 24)

Fourth, the issue has to do with the variance-bias trade off of the resampling studies. A low standard error for the adjustment of the (marginal) observed-score distributions does not guarantee a successful equating; it only tells us how stable the

result is (van der Linden, 1997). The SEE obtained by means of the resampling technique improves the reliability of the score transformation but this does not necessarily mean that the transformed scores are accurate or valid. The resampling technique makes the random error due to sampling and other sources of randomness stable because of the large number of iterations typical of these studies but this condition in and of itself is a only a necessary condition and by no means sufficient to establish the validity of the equated scores. In other words resampling improves the precision with which the equated scores are estimated but not necessarily their accuracy. Accuracy is a measure of how close the transformed scores are to the “true” scores. In other words, accuracy and bias may be considered synonymous. More attention needs to be paid to the validity of these transformed scores because ultimately these are the scores of interest.

Studies in test equating has a long history of focusing on SEE at the expense of bias, but indeed one of the main reasons behind the differences among equating methods is bias. Everything being equal (same test length, testing conditions, parallel forms, sample size, test design, and so forth) the only difference among different equating methods are their assumptions and the extent to which these are met or violated.

To address these issues related to accuracy, rather than using a raw score point transformation from the old form to the new form a different approach is used to compute the discrepancy in the score transformation. It is based on examinee’s abilities thereby taking into account possible measurement error that would otherwise be ignored by the traditional approach.

With respect to the equating methods under the NEAT design, there is wide agreement that post stratification equating (PSE) methods have a superior theoretical underpinnings compare to chained equipercentile (CE) methods but PSE are computationally more intensive than CE and there is a tendency for practitioners to shy away from CE methods because they appear too simple to be right (Sinharay & Holland, 2010a).

The findings of studies by Marco, Petersen and Stewart (1983), Livingston, Dorans and Wright (1990), Wang et al. (2008), Sinharay and Holland (2007) comparing the merits of methods under these two approaches in the non linear contexts and by Livingston (1990) and Wang, Lee, Kolen and Brennan (2006) in both linear and non linear contexts indicate that the CE methods tend to produce less equating bias but more random equating error than PSE methods when the new form population and the old form population differ substantially on the anchor test (Puhan, 2010). However, Harris and Kolen (1990) suggested using the PSE methods when groups differ in ability because of their superior theoretical underpinnings, but Marco et.al (1983), Livingston et.al (1990) concluded that the CE methods are superior when there are large group differences in ability.

Livingston (2004) provided a possible explanation of these observations when he demonstrated graphically that if the correlation between the total test and anchor test is not high, the PSE methods “over-adjust” the differences in form difficulty to the extent that the groups taking the alternate forms may appear to be similar in ability even though they actually are different. This explanation seems to support the findings of the previous

studies on this topic, i.e., the PSE methods tend to produce large amount of bias when the two equating samples differ in ability and the correlations between the anchor scores and the total test scores are weak. The CE methods on the other hand seems to be more robust to this kind of bias and tends to be less susceptible to increases in the size of the correlation between the anchor and the total test mainly because of their scaling procedure and symmetric nature.

Conclusion

The main conclusion drawn from the above review clearly points to the difficulty associated with equating small samples. Several considerations have to be made and these include the type of data collection design, the equating method used, the sample size, the difference in difficulty between the forms being equated, the choice of the criterion equating transformation or the definition of accuracy. Other considerations such as sample size, the range of the score scale relative to the number of examinees taking the test, the distribution of the scores on the test forms, whether we should select different methods for different ranges on the score scale all add to the complexity and difficulty in deciding whether to equate or not and whether a chosen method is the most appropriate one for the task at hand. With these issues in mind the decisions we make about each of them can have serious consequences of the test takers. For lack of a better method, identity equating and mean equating have remained the norm among practitioners and it is only in the last two years that the topic of small sample equating seemed to have regained some attention.

CHAPTER III

METHODOLOGY

The purpose of this study is to better understand how the accuracy of equating is affected by various test design and sampling factors. This chapter describes the design of a large-scale simulation study to investigate the impact of sample size, test length, item parameters, and equating method on equating error. To the extent that we are interested in investigating the accuracy of different equating methods under a wide variety of conditions, simulations are appropriate where the true parameters of interest are known and the quality of the results can be evaluated in an absolute sense. In addition, different measurement conditions (both realistic and perhaps extreme) can be designed into the simulations, making it possible to investigate the limits and conditions under which some procedures work or fail. Learning about conditions that contribute to failure in simulation studies is informative and useful; in practice, those same failures can be disastrous. Simulations also make it possible to examine a wide array of conditions which would be prohibitively costly with live subjects, while avoiding many of the issues related to extraneous or complex confounding factors that often arise with real data. Finally simulations are easy to conduct, can be replicated consistently and can be used for generalizing results about a limited selection of phenomena of interest.

However, we should be mindful that running simulations to test an array of scenarios will never fully characterize the true complexity of real data. As such any

conclusions and generalizations resulting from the simulations described here should not be construed as definitive conclusions.

The following subsections describe software resources employed to carry out all of the analyses, test characteristic and sampling conditions investigated, and the overall design of the large simulation study and replications. In addition, analysis methods are discussed leading the results reported in Chapter IV.

Software Resources Employed

The generation of examinee response data for this study is based on the three-parameter logistic (3PL) model derived from Item Response Theory (IRT). The program GEN3PLDATA (Luecht, 2007) is used to generate dichotomous (0, 1) response data under the three-parameter logistic (3PL) IRT model. The program is capable of simulating responses for virtually an unlimited number of examinees having abilities drawn from a user-defined normal distribution – i.e., $\theta \sim N(\mu, \sigma)$, and for tests having as many as 1,000 items.

The following three R packages were employed for various purposes: (a) sn- version 0.14; (b) equate-version 1.0-0; and (c) ggplot2- version 0.8.9. The “sn” package was used to generate the 3PL item parameters for test forms that were supplied to GEN3PLDATA. The “equate” package was used to conduct the observed-score equating transformations of the PSE class of methods for the NEAT designs. Only the Tucker, Levine Observed Score and Frequency Estimation methods from that package were used in this study. For the remaining equating methods, Identity, Circle-Arc, Chained Linear

and Chained Equipercentile routines were coded in R by the author to perform the necessary equating transformations. The ggplot2 package was used to plot all of the graphics used in this study.

To ensure that the R code performed as expected, several small verification studies were carried out using ACT mathematics test data presented in Kolen and Brennan (2004, chap.4). The Smoothed Chained Equipercentile (SCEE) and Smoothed Frequency Estimation (SFRE) reproduced exactly the same equated scores reported by Wang (2009). The CLIN, LLIN and TLIN methods were checked for accuracy against the results supplied in the “equate package” available for R.

The Circle-Arc (CARC) function could not be independently validated due to a lack of any published raw data validation sets or results. However, the programs for the Circle-Arc method were written according to the formulation of the Simplified Circle-Arc and easy-to-verify computations were independently checked for accuracy. For more detailed explanations and the technical aspects of the method, assumptions and computational formulas, see Livingston and Kim (2009).

The CARC method essentially constrains the estimated, equated scoring function curve to pass through two specified end points (i.e., forming an arc), with an empirically determined middle point. These three data points are all that are needed to perform the equating transformation of the alternate forms to the base forms. They are easily determined from the score distributions of the alternate form and the base form. One end point of the arc, called the upper end point, is determined by the maximum possible score on the test forms. In this study the upper end point used was 30 for the short forms and 60

for the long ones. The second point, called the lower end point, is generally determined as the scores that would be obtained by chance guessing. The same guessing proportion was applied to determine the lower end point of the CARC as was used in the IRT model for data generation (i.e., $Y_{min}=cn$, where c is the lower asymptote under chance guessing and n is the test length). For the 30-item test the lower end point was set at 5 and for the 60-item test the lower endpoint was set at 9, since c was set equal to .15 for all items used in this study. The middle point is empirically determined and was set at the same score that the mean score of the examinees of the alternate form group would be equated to by the chained linear method. All scores below the lower end point on the alternate form were equated to the same value on the base form. In other words scores less than 5 in the distribution of the new form assumed the same value on the base form for the short version. The same rule applied to scores that were 9 points or lower on the longer version.

Smoothing Techniques

Smoothing is typically performed to eliminate “jaggedness” in a mathematical function. In equating contexts, smoothing is typically applied to the score distributions of interest and the smoothed distributions are then equated to one another. Optionally, post-equating smoothing is also sometimes performed. Three factors were considered for purposes of determining how smoothing should be applied in this study. First, it was recognized that small-sample sizes used in this study could lead to nontrivial smoothing errors due to the choice of kernel function, bandwidth and other factors. In fact, under

the small-sample paradigm, there could be interactions and a propagation of errors due sampling, measurement, and smoothing that might ultimately create worse problems than any modeling or estimation bias introduced by particular equating methods. Second, equipercentile and frequency estimation would logically be susceptible to smoothing of relatively unstable percentile ranks, due to the small samples. This could result in large smoothing errors at certain points of the scale. This issue of percentile rank stability sometimes arises even when the sample size is quite large (Kolen and Brennan, 2004). Third, the problems associated with estimating a particular equating transformation function may be compounded when the sample size is small because of a greater likelihood of gaps in the score distributions due to the sparseness of data at some score points on the observed score scale. Although there are recognized statistical methods of collapsing score intervals and/or using unequally sized intervals for describing a relative frequency distribution of scores, most of those methods have not been vetted for equating applications, even with large samples.

There are many smoothing techniques that are available in the statistical graphics and mathematical literature, but in classical test equating, the two most commonly used techniques reported in the literature are the cubic-spline post smoothing (Kolen,1984) and the polynomial log linear smoothing (Holland and Thayer, 1987, 2000). To provide a consistent means of dealing with these smoothing issues, the polynomial log-linear smoothing technique was selected for this study. One reason for selecting log-linear smoothing was software availability, especially given the large number of analyses and replications that needed to be done. In this study, the log-linear smoothing capabilities of

the “equate” package in R were used to presmooth the scores of the alternate forms before they were equated by the chained equipercentile and frequency estimation equipercentile methods. The first three moments (mean, standard deviation, and skewness) and cross products of the score distributions of the anchor test and total test are constrained to remain the same by the log-linear smoothing. In other words, the estimated shape parameters associated with the joint bivariate score distribution of the base and alternate forms and with the marginal distributions are preserved, providing essentially the same source data (the smoothed distributions) for the chained equipercentile and the frequency estimation equating methods.

Data Generation and Study Factors

The main factors examined in this study included sampling characteristics and statistical test characteristics related to test score quality. Sampling characteristics include two key factors: (1) sample size and (2) group differences insofar as their score distributions. Test characteristics include three key factors: (1) test-form differences based on average item difficulty; (2) the ratio of measurement precision (i.e., test information) on the alternate form to the base form near the center of each sample; and (3) test length.

The three-parameter logistic (3PL) IRT model was used for all data generation:

$$\Pr u_i = 1 | \theta; a_i, b_i, c_i \equiv P_i \quad \theta = P_i = c_i + \frac{1 - c_i}{1 + \exp[-Da_i (\theta - b_i)]} \quad (3.1)$$

where θ is the examinee's proficiency score, D is a scaling constant, usually set to 1.7 to approximate a cumulative normal probability density function for representing the likelihood of correct responses to the items, a_i is an item discrimination parameter related to the degree to which an item contributes to measurement precision, b_i is an item difficulty (location parameter), c_i is a lower asymptote parameter typically associated with random noise or guessing near the lower regions of the score scale—especially for more difficult items. In this study, a constant value was used: $c_i=c=.15$. As noted in the Introduction, the choice to use 3PL IRT data generation was expedient insofar as providing a direct means of controlling various sampling and test characteristics. It also provided a very direct means of comparing outcomes in an absolute sense.

Under IRT, a test characteristic function (TCF) is computed as the conditional sum of the item response functions; that is, $TCF \equiv T(\theta) = \sum_i P_i(\theta)$, for $i=1, \dots, n$ test items. The TCF therefore represents a “true score” on any test where the 3PL item parameters and θ are known. Although neither θ nor the 3PL item parameters are known in practice, under an IRT simulation paradigm we actually do know “truth” and all functions of that truth. This provided a straight-forward way to compare equated observed scores on the alternate forms of the test (equated or X to the base form, Y) to TCF_Y and evaluate accuracy in an absolute sense. This IRT data generation and analysis approach does not appear to have been used in previous equating studies—almost assuredly not in the context of small-sampling equating research. Ultimately, having an absolute residual error—that is the difference between the TCF and the equated observed score—made it possible to focus on the conditions that increased either bias, as the summed aggregation

Table 3.1: Conditions Included in this Study

Equating Methods	(IDEN, CARC, CLIN, TLIN, LLIN, SFRE, SCEE)
Sample Sizes	$N=(25, 50, 100, 200, 400)$
Magnitudes of Group Separation	$STD=\Delta=\mu_{\theta(X)}-\mu_{\theta(Y)}\sim N(\Delta=0, .05, .10, .25; \sigma=1)$
Test Difficulty Differences	$SMD=\delta=\mu_{b(X)}-\mu_{b(Y)}\sim N(\delta=0, -.10, .25, -.50, -.75; \sigma=1)$
Test Lengths	$n=(30, 60)$
Average Item Discriminations ¹	$\mu_{a(X)}=(0.3, 0.6, 0.9)$ and $(0.5, 1.0, 1.5)$, $sd=0$
Base Form	(30 items, average item discrimination=0.6, Test Difficulty=0) (30 items, average item discrimination=1.0, Test Difficulty=0) (60 items, average item discrimination=0.6, Test Difficulty=0) (60 items, average item discrimination=1.0, Test Difficulty=0)
Base Form Group	$N=5,000$ examinees, $\theta\sim N(\mu=0, \sigma=1)$

¹ Note: These are the average item discriminations for the alternate forms. The set (0.3, 0.6, 0.9) is equated to a base form with average item discrimination of 0.6 whereas the set (0.5, 1.0, 1.5) is equated to a base form with average item discrimination of 1.0 such that the ratio of the average item discrimination of the alternate form to the base form remains at 0.5, 1.0 and 1.5 for each set.

of the signed residuals, or that increased the error variance (i.e., the summed aggregation of the *squared* residuals). Table 3.1 shows various factors and the different levels investigated for each factor.

Sample Size (N)

Based on the literature review, sample sizes ranging from 25 to 200 are common in small-sample equating studies. In this study, the sample sizes used are: 25, 50, 100, 200 and 400. Three reasons dictated the choice for including a sample size as large as 400. First, it provided a plausible, moderate-sized sample benchmark against which the

accuracy of other, smaller sample-size results could be compared. That is, it seems likely that a sample size of 400 might begin to mimic characteristics of the intended population score distribution, without resorting to a large-sample baseline comprised of thousands of examinees for each of the conditions. Second, it seemed very likely that this sample size of $N=400$ might be somewhat robust with respect to data sparseness in the score distributions. Third, it provided a reasonable gap to the next-in-line sample size ($N=200$) was reasonably small to chart any trends in the results, without resorting to transformations (e.g., logarithms).

Test Length (n) and Anchor Items

The impact of test length has been extensively studied in the literature and there is unanimous agreement that it has a direct effect on the reliability of test scores (e.g., Allen & Yen, 2002). By extension, it seems reasonable to assert that the magnitude of measurement errors will also have a direct impact on the accuracy of any equating method applied to those scores, especially when fallible observed scores are used in the equating process. This is particularly pertinent to equating methods falling under the umbrella of non-equivalent groups anchor test (NEAT) designs.

NEAT equating methods typically takes advantage of the variances and correlation between the base-form and anchor-test scores in their equating transformation functions. As the reliability decreases, the variances of observed scores proportionally increase and the correlations of observed scores are attenuated. Where distributions of the observed scores are directly manipulated (e.g., for equipercentile equating) or moments of the observed-score distributions are used to estimate equating transformation constants

(e.g., linear equating), the impact of differential reliability can be somewhat unpredictable.

In this study, a short test was operationally defined as consisting of 30 items and a long test consisted of 60 items. The number of common anchor items was held constant at 30 percent of the total test length. Thus, there were 9 common anchor items on the 30-item test and 18 common items on the 60-item tests. All the test forms assumed that the common anchor items were internal—that is, also counted in scoring. In contrast, external anchor items are sometimes used for equating purposes, but are not counted in scoring the examinees. In practice, the choice between using internal and external anchors—or some mixture of both—is largely a policy decision that is made by the testing agency.

Controlling Measurement Precision

As noted in the prior section, test length is perhaps the most common way of controlling measurement precision. Longer tests tend to be more precise than shorter tests. In this study, in addition to test length, score reliability was specifically controlled by manipulating the moments of two parameters in the IRT 3PL model: (1) the location of the items on each test form and (2) the amount of measure precision inherent in each test form. Under the IRT 3PL framework used for the simulations, the location of items is controlled by the item difficulty or b -parameters. The amount of measurement precision is affected by the spread of the b -parameters, but is primarily controlled by the mean of the item discrimination or a -parameters. The c -parameters—representing

chance noise or guessing in the response function— tend to reduce measurement precision.

In IRT, measurement precision is usually discussed in terms of a function known as the test information function (TIF). This function can be expressed as

$$TIF = I(\theta) = \sum_{i=1}^n I_i(\theta) = \sum_{i=1}^n \frac{D^2 a_i^2 (1 - P_i) P_i - c_i^2}{P_i (1 - c_i)^2}, \quad (3.2)$$

making use of Equation 3.1 (the IRT 3PL probability-response function). The TIF can be graphed to indicate the amount of measurement precision along the score scale, θ . The function is also inversely proportional to measurement error variance. The TIF provides a clear advantage over classical test theory reliability coefficients because it locates the precision relative to key decision points along the score scale and/or relative to the center of the distribution of examinee scores. In short, it was possible to directly manipulate not only the amount of measurement precision contributed by differential test length (i.e., the sum of item information), but also the psychometric quality of the test forms (mean of the a -parameters and mean and standard deviation of the b -parameters) used in the equating process.

Base Form Characteristics

Test equating requires one form to be used as the baseline or benchmark s test form—sometimes called the base-form scale. All subsequent, alternate test forms are equated to that base form scale. In theory, after equating, we assume that scores on the alternate or new test forms are psychometrically equivalent (i.e., exchangeable).

Therefore, equating estimates how an examinee would have been likely to perform on the base form, even though he or she took an alternative form of the test. This leads to the statement that, after equating, examinees ought to be indifferent as to which forms of the test they took. The equated scores are randomly equivalent to the base form and have the same content-referenced or normative interpretation as scores on the base form.

The base form (denoted here as Form *Y*) is sometimes also called the “old form” or the “reference form”. The alternate or to-be-equated form (denoted as Form *X*) is referred to as the new form. In this study four old forms were created to simulate four distinct base form characteristics: (1) 30 items (short test) with $\mu(a)=.6$ (low discrimination); (2) 30 items with $\mu(a)=1.0$ (good discrimination); (3) 60 items (long test) with $\mu(a)=.6$; and (4) 60 items with $\mu(a)=1.0$.

Each of these base-form conditions actually implies a separate study (numbered 1 to 4) where the conditions affecting the alternate forms are considered within the context of a base form with different measurement characteristics. Thus, for Studies 1 and 3, the base forms are not as discriminating as in Studies 2 and 4. Studies 1 and 2 employ relatively short base-form tests, while Studies 3 and 4 use longer base-form tests. The alternate forms were constructed to have the same test length as the base form in each of the studies. Therefore, the issue of equating alternate forms of lengths different than the base form is not considered in this study. However, it is worth noting that the manipulation of the mean of the item difficulty and discrimination parameters under the IRT framework indirectly alters the reliability of the base and alternate forms, producing scenarios where unequally reliable tests are being equated.

Unique and Anchor Items

All test forms consist of a set of unique operation items and a set of common, internal anchor items. The shorter forms in Studies 1 and 2 consist of 30 items made up of 21 unique items and nine anchor items. The difficulty of the anchor items ranges from -1.6 through 1.6 and increases with a step size of 0.4, producing a uniform distribution of item difficulty for the anchor items

The forms in Studies 3 and 4 are twice the length of those in Studies 1 and 2. They have 42 unique items and 18 anchor items. The same uniform distribution of item difficulty parameters is applied to these longer tests by doubling the number of anchor items at each value from -1.6 to 1.6.

The remaining unique items are generated so that the difficulty of the total test follows a standard normal distribution, with the mean determined by the condition specified in Table 3.1: $SMD \sim N[\mu(b) = (0, -.10, .25, -.50, -.75), \sigma(b) = 1]$. Furthermore the item discrimination parameters (both for the anchor and unique items) on all forms are set to the fixed values of $a = \mu(a) = .6$ for Studies 1 and 3 and $a = \mu(a) = 1.0$ for Studies 2 and 4, with a variance of $\sigma(a) = 0.0$. These somewhat unrealistic scenarios of having all items on a particular test form have the same discrimination parameters nonetheless helped isolate the contribution of differential discrimination between the base and alternate forms.

Magnitude of Group Separation (STD)

The magnitude of group separation or group effect can be conceptualized as the ability differences between the groups taking the alternate form relative to the base form. In this study, all examinees taking the base forms are sampled from a standard normal

distribution with a mean of zero and standard deviation of one. In addition, four distinctly different populations of examinees take the alternate forms.

Groups in the first category are sampled from the same standard normal distribution as the base form groups; that is, $[\mu(\theta)=0, \sigma=1.0]$. Groups in the second category are slightly more able and are sampled from a distribution with a mean ability $1/20^{\text{th}}$ of a standard deviation unit higher than the base-form group but with the same variability as the base form groups—i.e., $[\mu(\theta)=.05, \sigma=1.0]$. Groups in the third category are even more able and are sampled from a distribution with a $\mu(\theta) = 0.1$ higher and a standard deviation of 1.0. Finally, examinees in the fourth category were sampled from a very able population of examinees, with abilities on average 0.25 standard deviation units higher than the base form groups, that is, $\theta \sim N[\mu(\theta)=0.25, \sigma(\theta)=1]$. It is worth noting that in test equating, mean differences between 0.05 and 0.1 are generally considered relatively large, whereas a mean difference of 0.25 is usually considered to be an extremely large difference (Wang et al. 2008). The inclusion of a difference of only $\Delta=.05$ in this study helps to test the assertion that Wang et al. makes about the impact of this very small difference relative to no group difference (i.e., random sampling from the same population to obtain randomly equivalent groups) on equating accuracy.

Summary of Conditions for the Generation of Alternate Forms

Differences in test form difficulty were modeled by an effect size representing the difference in the mean difficulty of the base and alternate form: $SMD=\delta=(0, -0.1, -0.5,$

-0.75, 0.25). Each of these test difficulty effects are crossed with three discrimination ratios representing the ratio of the mean item discrimination of the alternate form to the mean item discrimination of the base form. These ratios are $R=\mu(a_{alt})/\mu(a_{base})=(0.5, 1.0, 1.5)$. For example, with a base-form discrimination of 0.6, $R=0.5$ indicates an average a -parameter of 0.3 for the alternate forms. Obviously, that level of discrimination would result in a very poor quality, psychometrically speaking. However, that is one of the advantages of simulations, as noted earlier. It allows us to explore the impact of otherwise unacceptable conditions of measurement on the results, without subjecting actual examinees to those conditions.

This simulation design generates five test difficulty effects levels crossed with three discrimination ratios for a total of 15 different alternate forms for each of the four base form studies. Therefore, in reality, there are 60 distinct cells in the simulation study design matrix for test characteristics, alone. However, those 60 test characteristic conditions are then crossed with five sample sizes and four group mean difference conditions (STD), creating 1,200 total conditions. Finally, each of those conditions is replicated ten times to provide sampling distributions of the results, and each data set is then analyzed using each of the seven small-sample equating methods to provide side-by-side comparisons (i.e., a total of 84,000 separate equating analyses).

Data Generation Procedures

The many crossed simulation study design conditions require analyzing over 12,000 separate data sets per Study. A four-step process is employed to obtain those data

sets. The process generates a large data set for each study condition as the sampling frame, and then randomly selects the smaller samples, without replacement, from that sampling frame. Other equally effectively sampling strategies could, of course, be implemented.

In step one, item parameters are generated for each combination of test length, item difficulty difference, and ratio of the mean discrimination parameter for the base and alternative forms. This procedure was repeated 10 times for each condition to produce ten randomly parallel alternate forms for a particular set of design conditions. For example, ten randomly parallel alternate forms were generated for a test of length of 30 items, with an average item difficulty 0.25 and an average item discrimination of 0.9.

For the second step, the item parameters are input to GEN3PLDATA, which in turn produces the complete response data for 5,000 examinees—that is, the sampling frame. This process was further repeated for each of the four ability distributions—i.e., $\mu(\theta)=0.0$, $\mu(\theta)=0.05$, $\mu(\theta)=0.1$, or $\mu(\theta)=0.25$. A total of 600 population response data corresponding to ten alternate forms by four levels of STD by three levels of the discrimination parameters and five levels of SMD are generated for each of studies 1 to 4 (base form test lengths of 30 and 60 items by two levels of discrimination: 0.6 and 1.0). Each population response data represents one distinct alternate-form sampling frame.

In step 3, 5,000 examinee abilities were generated by sampling from a unit-normal distribution by GEN3PLDATA for each of the four base forms conditions (i.e., studies 1 to 4) corresponding to the test-length by level of base-form discrimination

conditions. Each base form was subsequently matched with the ten alternative test forms under a particular condition.

In the final step, the smaller data sets are sampled from each of the larger frames, with replacement. That is, samples of 25, 50, 100, 200, and 400 are drawn from each alternate-form sampling frame of 5,000 examinee records. Those smaller samples were then used in the equating steps, matching the smaller sample of data with the N=5,000 base-form sample for that study. (Note: there is no possible sampling overlap between the base-form and alternate data sets.)

It is important to realize that, in addition to generating the raw scores and corresponding total (number-correct) scores for each simulated examinee using the IRT simulation procedures outlined earlier, GEN3PLDATA also computes each examinee's true score—that is, the test characteristic function (TCF)—using the item parameters from the base form. Therefore, two scores exist for each simulated examinee: (1) the observed, to-be-equated number-correct score on the alternate form and (2) the TCF value or true score on the base form.

Equating Steps

Each alternate-form, small-sample data set is paired with the corresponding base form data set for each set of design conditions: sample size, group differences in ability, test length, test form difficulty differences, discrimination differences between and base form and alternate form, and discrimination of the base form, itself.

The observed total-test scores for each of the ten replications of samples of 25, 50, 100, 200 and 400 examinees taking the alternate (NEW) form are then equated seven times, each time using one of the seven equating methods (IDEN, CARC, CLIN, TLIN, LLIN, SFRE, SCEE). The samples were drawn with replacement and equated to the observed scores from the corresponding OLD form population. This process provides a form-specific equating function for converting the alternative-form scores to the base-form scale. As noted above, each simulated TCF or true score on the base-form scale can be compared in an absolute sense to each of the seven equated observed scores. A total of 21,000 equating transformations are conducted for each of the four studies study.

Evaluation of Equating Accuracy and Stability

Two approaches for evaluating equating accuracy in small samples are commonly used and reported in the literature: an analytical approach and a bootstrap method. The bootstrap resampling method can be applied to virtually all equating designs and methods whereas the analytical approach is limited to some methods and designs (Wang, 2009; Wang et al., 2008). In particular, most studies report equating error in terms of the standard error of equating (SEE), bias, and a root mean square error (RMSE), conditional on the base-form scores. When real data are used, “truth” must be estimated by either resampling and computing a mean (expected value), or using an extremely large sample of examinee data—if it even exists—to approximate the “correct” results against which to compare the equated scores. In most cases the accepted “gold” standard is to use the equipercentile equating method or some version of an observed-score or estimated true-

score IRT equating method such as Stocking and Lord (1983) to establish the criterion for evaluating accuracy and stability of the equating functions. It is not unusual to find similar approaches adopted even in the case of simulation studies (Wang et.al, 2008; Sinharay & Holland, 2010a, 2010b), despite the fact that the “true” scores on the base form could have been determined in studies of that nature.

When large samples are used, these expectation-based methods to establish the criterion scores are probably reasonable—especially when real data are used in the study (Harris & Crouse, 1993). However, using these large sample criterion approaches to evaluate the accuracy of an equating method may not be appropriate in particular when the samples are small and “true” scores are unknown (e.g., for small testing programs that do not have the luxury of large-sample, well-behaved data from which to mimic small sampling conditions).

Using the IRT-based TCF-scores, using the base-form item characteristics, this study employs a very direct approach for determining the absolute error of equating at the level of individual [simulated] examinees, consistent with IRT simulation studies in general with the approach to evaluating equating suggested by van der Linden (2006) and implemented by Wang, Lee, Brennan & Kolen, 2008).

The residual of interest is $d_j = X_j^* - Y_j$, where $Y_j = T_X(\theta_j) = \sum_i P_i(\theta_j)$, as the true score for the $i=1, \dots, n$ test items on the base form and X^* is the equated alternate-form score. Averaging d_j (signed) provides a convenient index of BIAS. That is,

$$BIAS_{eq_x} = \frac{1}{N} \sum_{j=1}^N d_j \quad (3.3)$$

The conditional or unconditional expectation of BIAS is zero. Averaging d_j^2 and taking the square root provides a root mean squared difference (RMSD) statistic, representing the error variability:

$$RMSD_{eq_x} = \sqrt{\frac{1}{N} \sum_{j=1}^N d_j^2} \quad (3.4)$$

Each condition in this study provides a natural way to breakdown the results. In addition, having the ten replications of every data set provides a small sampling distribution of those statistics (i.e., minimum and maximum values, means, and standard deviations of BIAS and RMSE). Given the large number of potential multi-way interactions among the study design conditions, it was decided to focus on reporting primarily first-order effects in terms of results. That decision was pragmatic in nature, especially insofar as providing a useful level of interpretation of the results in Chapter IV.

Preliminary Analysis of Base Forms

Initially, the response data for the four base forms (two test lengths by two levels of average item discrimination) were generated and the corresponding score distributions were analyzed to ensure that their characteristics adequately represented the intended population characteristics. These preliminary, screening analyses were essential because there is only one distribution of base form scores in each study and that same distribution

is paired with all of the alternate forms (each, in turn, equated under all seven equating methods). A base form producing aberrant response data and scores under the 3PL model

Table 3.2: Summary Statistics for Base Form Observed Total, True Total Scores and Anchor Scores in the Population

Base Form (Study)	Score	Mean	SD	Median	Range	Skew	Anchor/Total Correlation	Reliability
1	Obs.Total	17.278	5.270	17	29	-0.073	.814	.782
	True.Total	17.281	4.661	17.175	22.164	0.018		
	Anchor	5.165	1.914	5	9	-0.134		
2	Obs.Total	17.191	6.128	17	29	-0.019	.875	.870
	True.Total	17.173	5.716	17.000	24.495	0.043		
	Anchor	5.203	2.060	5	9	-0.135		
3	Obs.Total	34.548	9.922	35	53	-0.058	.882	.875
	True.Total	34.512	9.281	34.543	44.090	-0.004		
	Anchor	10.378	3.288	10	18	-0.104		
4	Obs.Total	34.263	11.960	34	57	0.054	.923	.938
	True.Total	34.273	11.586	34.056	49.183	0.079		
	Anchor	10.337	3.773	10	18	-0.060		

Table 3.3: Summary Statistics of Ability Distributions on Base Form Populations

Base Form	Mean	SD	Median	Min	Max	Range	Skew
1	0.000	1.000	-0.027	-3.012	2.998	6.010	0.053
2	0.000	1.000	-0.003	-2.933	2.953	5.886	-0.036
3	0.000	1.000	0.001	-2.992	2.950	5.942	0.003
4	0.000	1.000	0.001	-3.063	3.040	6.102	0.046

would potentially jeopardize the interpretation of the subsequent results. Table 3.2 shows the distribution of the observed total test scores, the true total test (TCF) scores, and the observed scores for the anchor item sets for Studies 1 to 4 in the full population of 5,000 examinees.

Table 3.4: Anchor/Total Correlation and Reliability of Alternate Forms Summarized by Average Item Discrimination and Test Length for each Study

Study	Length	Discrimination	Anchor/Total Correlation		Reliability	
			Min	Max	Min	Max
1	30	0.3	0.68	0.73	0.54	0.58
		0.6	0.80	0.83	0.76	0.81
		0.9	0.81	0.87	0.77	0.89
2	30	0.5	0.77	0.81	0.71	0.77
		1.0	0.86	0.88	0.86	0.91
		1.5	0.86	0.91	0.86	0.93
3	60	0.3	0.76	0.79	0.69	0.74
		0.6	0.88	0.89	0.86	0.91
		0.9	0.88	0.92	0.86	0.94
4	60	0.5	0.85	0.87	0.82	0.88
		1.0	0.91	0.93	0.92	0.95
		1.5	0.91	0.95	0.93	0.97

From Table 3.2 it should be obvious that the Study 1 test characteristics are mediocre compared to the other base forms. That is, the Study 1 base-form test (30 items with an average discrimination of .6) had only marginal reliability and correspondingly reduced variance of the observed scores. It seems predictable that the results might be inconsistent for some conditions when these forms are equated to one another. In fact, all of the test forms with mean discrimination of only 0.3 could be predicted to result in very inaccurate and inconsistent equated scores.

Table 3.3 provides a summary of the population ability distributions generated for each of the 4 four base forms to which samples taking the alternate forms are equated to. The statistics clearly show that the ability distributions followed the intended unit-normal distributions.

Preliminary analysis was also conducted to verify determine the test reliability and the correlation of anchor to total observed scores for the 600 (5 SMD by 3 a-ratio by 4 STD by 10 replications) population data generated per study. Table 3.4 shows the range of these statistics summarized over 200 (4 STD by 5 SMD by 10 replications) conditions associated with each level of discrimination. No data is presented for summaries by test difficulty or the magnitude of group separation because they did not affect the correlation or reliability.

CHAPTER IV

RESULTS

To present the results of this study, a graphical approach is used to support the data from the corresponding tables and portray the trends and patterns of the equating error incurred by the different methods across the various conditions without compromising the fidelity and accuracy of the results. The results relating to the two measures of accuracy av.BIAS and av.RMSD are presented separately with respect to the factors studied. Recall that the factors examined in this study are sample size, the magnitude of group separation, the magnitude of the difference in test difficulty, the test length, and the ratio of the average item discrimination of the alternate forms to the base form. The findings relating to the first research question about the effect of each of these factors on the equating BIAS and the associated RMSD for the various methods are presented in the order the listed below:

Effect of Sample Size on Average BIAS

Effect of Sample Size on Average RMSD

Effect of Magnitude of Group Separation on Average BIAS

Effect of Magnitude of Group Separation on Average RMSD

Effect of Magnitude of Differences in Form Difficulty on Average BIAS

Effect of Magnitude of Differences in Form Difficulty on Average RMSD

Effect of Discrimination Ratio on Average BIAS

Effect of Discrimination Ratio on Average RMSD

Effect of Test Length on Average BIAS

Effect of Test Length on Average RMSD

In addressing research question 1, the emphasis is placed on the impact of the factor of interest on the accuracy and consistency of the equating results. Because research question 2 addresses the impact of the equating conditions on the interchangeability of the equated scores among the equating method and uses the same data tables and charts as question 1, it is answered concurrently with question 1 under the same heading listed above. The connections between these two questions are very strong even if they address two different issues.

In addition, because the study includes four separate smaller studies the findings about the impact that the base form characteristics of each study have on the accuracy, stability of the equated scores and their interchangeability across the equating methods comparisons are also included. As much as this approach makes the analysis harder, it facilitates the examination and comparisons of the impact of the base form test length as well as its average item discrimination in the presence of the factor of interest. All the findings described for each sub question from both research questions 1 and 2 are derived from the same tables and charts. The third research question is based on the findings derived from research questions 1 and 2. As a result the set of rules to guide the choice of the most appropriate conditions and methods when equating can be considered to be successful or fail to work is presented in Chapter V following the summary of the results from this chapter.

Effect of Sample Size on Average BIAS

Question 1: Accuracy of Equating Methods and av.BIAS

This section describes the results of the average BIAS of the equating methods as a function of sample size for Studies 1-4. For each level of sample size the av.BIAS, computed over 10 replications of randomly parallel forms, are pooled over all combinations of SMD, a- ratio, STD and test length. Table 4.1 provides the summary statistics (av.BIAS and standard error of the av.BIAS) across the equating methods, conditioned on sample size for all four studies. Figure 4.1 shows the 95% CI of the av.BIAS of the various methods as a function of sample size for the various equating methods across the Studies 1-4. The columns of Figure 4.1 refer to the four studies.

The distributions in Figure 4.1 correspond to data in Table 4.1 within the range where the 95% Confidence interval (CI) lie between -.5 to .5. Several observations can be made from this figure. The first and most important result is that the av.BIAS of the equating methods is essentially independent of sample size. The bouncing patterns relative to the zero bias line of the location of the various equating methods as the sample size increases, illustrated in Figure 4.1, confirm this finding. However, except for the CARC and IDEN methods the CI's become progressively narrower with increasing sample size. The variances are generally more pronounced for sample sizes 25 and 50 and gradually tapers off at around size 200 for the 30 item tests and at around size 100 for the 60 item tests. These results suggest that even if the bias of the equating methods is invariant to sample size, chances are that equated values on a form may be systematically under or overestimated on any given equating when the sample size is small. In the

Table 4.1: Effect of Sample Size on Average BIAS and SE of Equating Methods for Studies1-4

<i>Study 1</i>	<i>Size</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	25	Mean	-0.11	-0.12	-0.18	-0.18	-0.03	0.04	0.90
		SE	0.04	0.04	0.04	0.04	0.04	0.06	0.22
	50	Mean	-0.14	-0.15	-0.21	-0.20	-0.06	0.03	0.90
		SE	0.03	0.03	0.03	0.03	0.03	0.05	0.22
	100	Mean	-0.11	-0.12	-0.18	-0.18	-0.02	0.05	0.88
		SE	0.03	0.03	0.03	0.03	0.03	0.05	0.22
	200	Mean	-0.09	-0.10	-0.17	-0.16	0.00	0.07	0.90
		SE	0.02	0.02	0.03	0.03	0.02	0.05	0.22
	400	Mean	-0.10	-0.11	-0.17	-0.17	-0.01	0.06	0.91
		SE	0.02	0.02	0.03	0.03	0.02	0.05	0.22
<i>Study 2</i>	<i>Size</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	25	Mean	-0.12	-0.14	-0.18	-0.17	-0.05	0.16	1.27
		SE	0.03	0.03	0.04	0.04	0.04	0.06	0.26
	50	Mean	-0.20	-0.22	-0.26	-0.25	-0.14	0.09	1.23
		SE	0.03	0.03	0.03	0.03	0.03	0.07	0.26
	100	Mean	-0.17	-0.18	-0.22	-0.21	-0.11	0.12	1.26
		SE	0.03	0.03	0.03	0.03	0.02	0.06	0.26
	200	Mean	-0.15	-0.17	-0.20	-0.19	-0.10	0.14	1.26
		SE	0.02	0.02	0.02	0.02	0.02	0.06	0.26
	400	Mean	-0.16	-0.17	-0.21	-0.20	-0.10	0.14	1.26
		SE	0.02	0.02	0.03	0.03	0.02	0.06	0.26
<i>Study 3</i>	<i>Size</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	25	Mean	-0.09	-0.10	-0.15	-0.14	-0.04	0.57	0.94
		SE	0.03	0.03	0.03	0.03	0.04	0.05	0.22
	50	Mean	-0.08	-0.09	-0.13	-0.13	-0.04	0.59	0.94
		SE	0.03	0.03	0.03	0.03	0.03	0.05	0.22
	100	Mean	-0.09	-0.10	-0.14	-0.13	-0.04	0.58	0.93
		SE	0.02	0.02	0.02	0.02	0.02	0.05	0.22
	200	Mean	-0.10	-0.11	-0.15	-0.15	-0.05	0.57	0.92
		SE	0.02	0.02	0.02	0.02	0.02	0.05	0.22
	400	Mean	-0.10	-0.11	-0.15	-0.14	-0.05	0.58	0.93
		SE	0.02	0.02	0.02	0.02	0.02	0.05	0.22
<i>Study 4</i>	<i>Size</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	25	Mean	-0.10	-0.12	-0.16	-0.14	-0.06	0.68	1.21
		SE	0.03	0.03	0.03	0.03	0.03	0.07	0.26
	50	Mean	-0.08	-0.10	-0.13	-0.12	-0.04	0.71	1.24
		SE	0.03	0.03	0.03	0.03	0.03	0.07	0.26
	100	Mean	-0.09	-0.11	-0.15	-0.13	-0.05	0.70	1.25
		SE	0.02	0.02	0.02	0.02	0.02	0.07	0.26
	200	Mean	-0.08	-0.10	-0.14	-0.12	-0.04	0.71	1.25
		SE	0.02	0.02	0.02	0.02	0.02	0.07	0.26
	400	Mean	-0.06	-0.07	-0.11	-0.10	-0.01	0.73	1.25
		SE	0.02	0.02	0.02	0.02	0.02	0.07	0.26

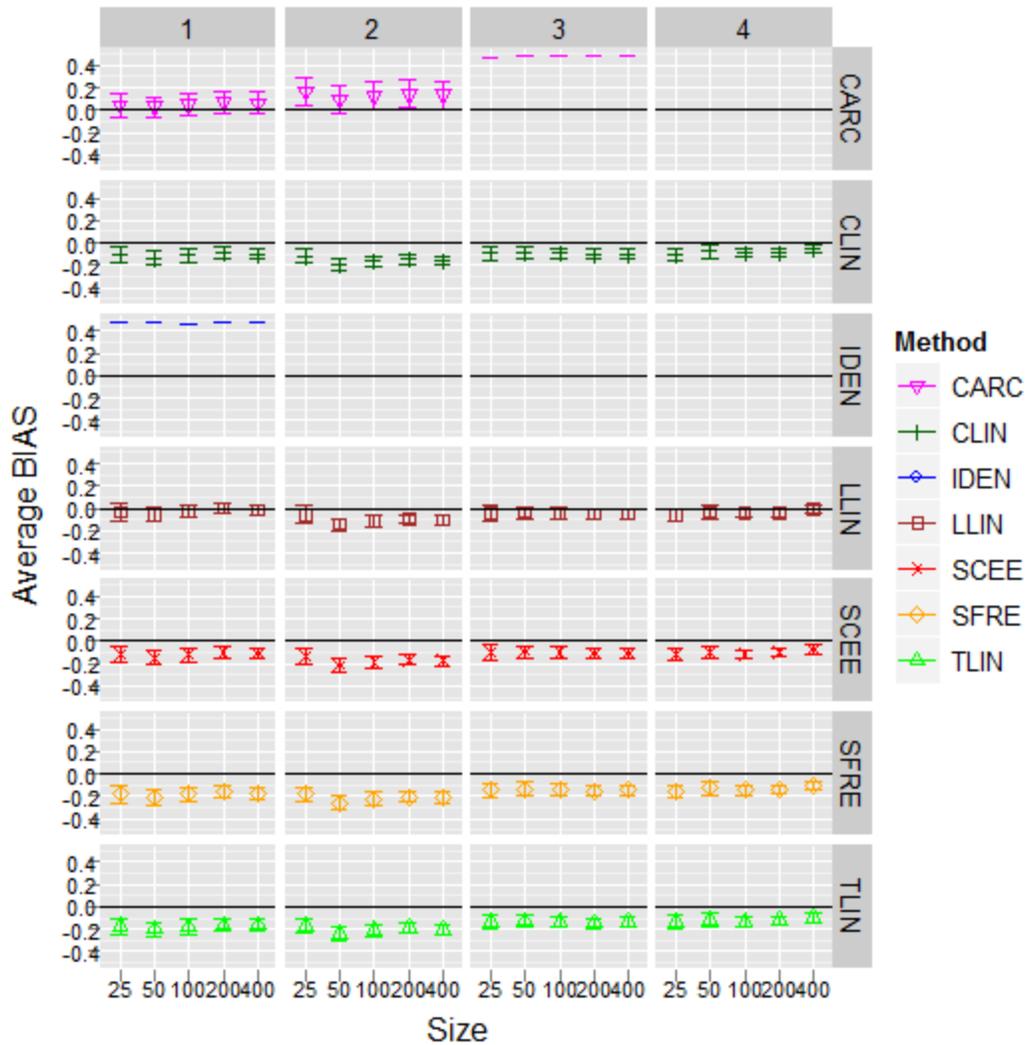


Figure 4.1: Distribution of the 95% CI of the Average BIAS as a Function of Sample Size for the Various Methods by Studies Combinations

universe of all possible conditions, equating very small samples is more likely to lead to less consistent equating results than larger sample size.

Except for the CARC method, examination of the trends across the four studies, show that there is no major improvement in the bias between studies 1 and 3, whereas there is a consequential reduction of the bias between Studies 2 and 4 for all the NEAT

but the LLIN method and sample size 25. In fact, the average BIAS of the NEAT equating methods in Study 4 is nearly almost half the value in Study 2. These results indicate that the base form conditions have a direct effect on the equating bias. The main conclusion that be drawn from these results is that the base form test length has no effect on the bias on the equating methods but the base form average item discrimination is a serious factor that has a direct influence on the bias of the equating methods, especially the theoretically related methods (CLIN and SCEE or SFRE and TLIN) of the NEAT group.

Figure 4.2 shows the 95 percent CI of the av. BIAS of the various methods for the various sample size by studies combinations. This figure is used here to compare the accuracy of the methods relative to one another. The LLIN method stands out as the least biased staying generally well-centered about or close to the zero bias line. The NEAT methods in general can be arranged in decreasing order of accuracy based on the average BIAS statistic as: LLIN, CLIN, SCEE, TLIN, SFRE. The LLIN methods stand out as distinctly different from the others and the SFRE is the most biased of them of all. Relative to one another, they share about the same degree of variability.

Examination of Figures 4.1 or 4.2 show that there is no plot for the data corresponding to the IDEN method because its average BIAS as shown in table 4.1 ranges from 0.88 to 1.27 which is well outside the range of the data in the plots. It is the most biased and most unstable among all the seven methods. The plots for the CARC appears only in Studies 1 and 2 because the range of its av. BIAS for the 60 item tests lies between 0.57 and 0.73 which is well beyond the limits in the charts. This is shown in

Figure 4.2, by the plots relating to Studies 1 and 2. Only in Study 1 does the CARC have a competing chance against the other NEAT methods, otherwise it can be considered as generally very biased. Its equating results are least biased when the base forms conditions are poor: (30 items, and average item discrimination of 0.6). When the base forms become more discriminating or the test length is doubled it becomes more biased.

Overall, sample size does not affect the bias of the equating methods. Arranged in order of decreasing accuracy the equating methods follow this sequence: LLIN, CLIN, SCEE, TLIN, SFRE, CARC and IDEN. The av.BIAS due to the NEAT methods has about the same variability but the IDEN and CARC are relatively more unstable.

Question 2: Interchangeability of Equating Results and av.BIAS

Figure 4.2 can also be used to compare the degree of exchangeability of the equating results among the various methods based on the av.BIAS. The most striking feature of Figure 4.2 is that the NEAT methods can be split into three very distinct groups: LLIN, CLIN/SCEE and TLIN/SFRE. The equating results of the methods within a cluster may be considered exchangeable with one another and depending on the sampling or test or base form conditions these clusters may join or deviate from one another to form new clusters.

It is clear from Figure 4.2 that the CLIN method and SCEE mimic each other very closely and can be considered to form one cluster. The SFRE and TLIN methods function in their own way and can be considered as another cluster. The formation of these clusters is consistent across the four studies and across all the levels of sample size. In other words their formation is robust to variations in sample size or the base forms

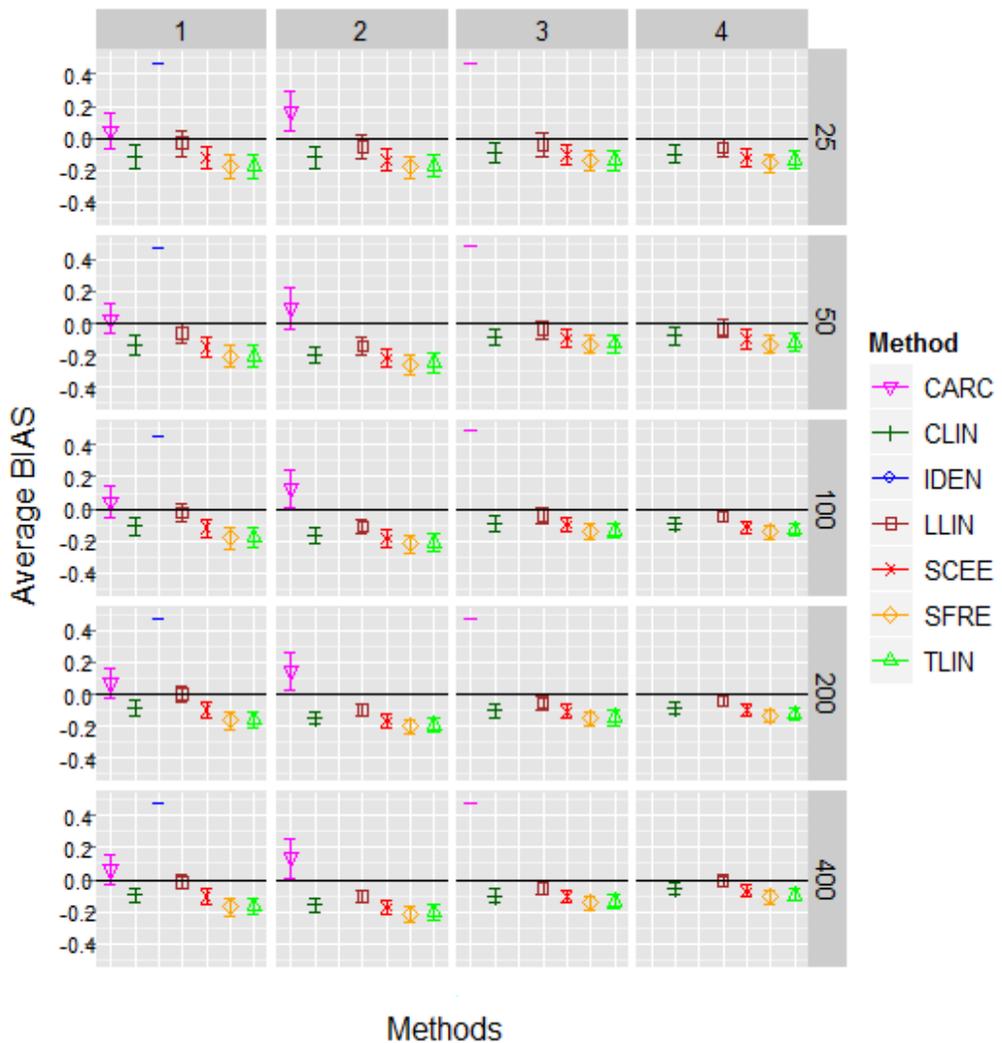


Figure 4.2: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various Sample Size by Studies Combinations

conditions. Visual inspection of Figure 4.2 reveals that the relative discrepancies among the equating methods remain almost constant across sample size. It seems safe to conclude that the degree of exchangeability (based on the overlap of the 95% CI) of the equating scores relative to the bias of the equating methods, in particular the NEAT methods, is independent of sample size.

Effect of Sample Size on Average RMSD

Question 1: Accuracy of Equating Methods and av. RMSD.

This section presents the results of the average RMSD of the various equating methods as a function of sample size for all four studies. For each level of sample size the av.RMSD, computed over 10 replications of randomly parallel forms, are pooled over all combinations of SMD, a-ratio, STD and test length. Table 4.2 shows the data and Figure 4.3 shows the distributions of the 95% confidence interval of the average RMSD of the various methods as a function of sample size, based on the results in Table 4.2. Figure 4.4 shows the 95% CI of the av.RMSD of the various methods after for the various sample size by studies combinations.

The data from Table 4.2 shows that within any one of Studies 1-4, there is a progressive but slow decline in the av.RMSD as sample size increases from 25 to 400 mainly for the NEAT design methods. For example for the CLIN method in Study1, the av.RMSD is highest (2.73) for sample size 25 and is smallest (2.64) for sample size 400. In Study2, the range is from 2.45 to 2.51 for the same method. The same trend (2.01 to 2.07 and 1.88 to 1.93) is maintained for Studies 3 and 4 respectively.

Based on Figure 4.3, the overlapping of the 95 percent CI intervals of the av.RMSD for any of the methods suggests that the drop in the total equating error from one level of the sample size to the next is not statistically large. The logical conclusion would be that sample size doesnot affect the av.RMSD , just as it had no effect on the av.BIAS in the section described above. However, unlike in the case of the av.BIAS which fluctuated in a haphazard way with sample size, in this case there is no bouncing

Table 4.2: Effect of Sample Size on Average RMSD and SE of Equating Methods for Studies1-4

<i>Study 1</i>	<i>Size</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	25	Mean	2.73	2.73	2.72	2.72	2.79	2.76	3.17
		SE	0.06	0.06	0.07	0.07	0.05	0.04	0.09
	50	Mean	2.70	2.69	2.69	2.71	2.73	2.76	3.19
		SE	0.07	0.07	0.08	0.07	0.06	0.05	0.09
	100	Mean	2.64	2.63	2.64	2.66	2.65	2.72	3.17
		SE	0.06	0.07	0.07	0.07	0.05	0.04	0.09
	200	Mean	2.65	2.63	2.64	2.67	2.64	2.73	3.19
		SE	0.06	0.07	0.07	0.07	0.05	0.04	0.09
	400	Mean	2.64	2.62	2.63	2.66	2.63	2.72	3.18
		SE	0.06	0.07	0.07	0.07	0.05	0.04	0.09
<i>Study 2</i>	<i>Size</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	25	Mean	2.51	2.46	2.45	2.51	2.56	2.55	3.26
		SE	0.06	0.06	0.06	0.06	0.05	0.04	0.13
	50	Mean	2.49	2.42	2.41	2.49	2.51	2.54	3.25
		SE	0.05	0.06	0.06	0.06	0.05	0.04	0.13
	100	Mean	2.47	2.40	2.41	2.48	2.48	2.54	3.27
		SE	0.06	0.06	0.06	0.06	0.05	0.04	0.13
	200	Mean	2.46	2.37	2.38	2.47	2.46	2.53	3.28
		SE	0.06	0.06	0.06	0.06	0.05	0.04	0.13
	400	Mean	2.45	2.36	2.36	2.46	2.44	2.52	3.27
		SE	0.05	0.06	0.06	0.06	0.05	0.04	0.13
<i>Study 3</i>	<i>Size</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	25	Mean	2.07	2.06	2.03	2.05	2.14	2.22	2.69
		SE	0.05	0.05	0.05	0.05	0.05	0.04	0.10
	50	Mean	2.03	2.01	1.98	2.01	2.09	2.20	2.67
		SE	0.05	0.05	0.06	0.05	0.05	0.05	0.10
	100	Mean	2.02	1.98	1.96	2.00	2.06	2.20	2.68
		SE	0.05	0.05	0.06	0.05	0.05	0.05	0.10
	200	Mean	2.01	1.97	1.95	1.99	2.04	2.19	2.68
		SE	0.05	0.05	0.06	0.05	0.05	0.04	0.10
	400	Mean	2.00	1.96	1.94	1.98	2.03	2.19	2.68
		SE	0.05	0.05	0.06	0.05	0.05	0.04	0.10
<i>Study 4</i>	<i>Size</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	25	Mean	1.93	1.84	1.81	1.91	1.97	2.07	2.82
		SE	0.05	0.05	0.05	0.05	0.05	0.05	0.15
	50	Mean	1.92	1.81	1.79	1.91	1.95	2.09	2.85
		SE	0.05	0.05	0.05	0.05	0.05	0.05	0.15
	100	Mean	1.90	1.78	1.76	1.89	1.93	2.09	2.86
		SE	0.05	0.05	0.05	0.05	0.05	0.05	0.15
	200	Mean	1.89	1.76	1.75	1.88	1.91	2.09	2.86
		SE	0.05	0.05	0.05	0.05	0.05	0.05	0.15
	400	Mean	1.88	1.75	1.74	1.87	1.90	2.10	2.86
		SE	0.05	0.05	0.05	0.05	0.05	0.05	0.15

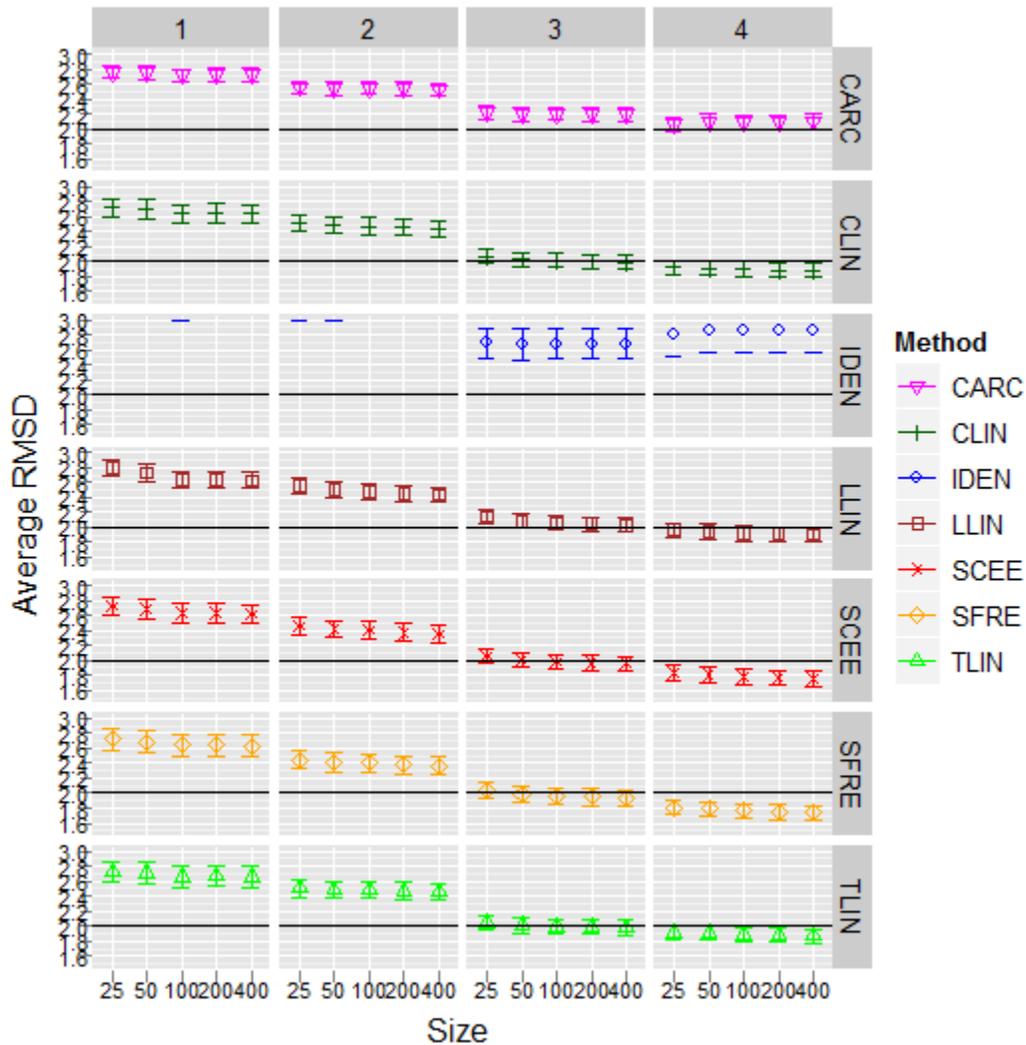


Figure 4.3: Distribution of the 95% CI of the Average RMSD as a Function of Sample Size for the Various Methods by Studies Combinations

but rather an unmistakable but slow and nearly asymptotic decline as the sample size increases from 25 to 400. Generally there is a larger dip in the error from sample size 25 to 50 and by sample size 100, the change is nearly asymptotic. The one exception to this rule is Study 4, where there is still the slow decline but the drop from size 25 to size 50 is not as pronounced as in the other studies. One potential explanation for this is that the

base from condition could have been the cause behind this exception.

In terms of stability of the average RMSD, the SE across the five levels of sample size remained fairly constant for most methods within a given Study. In other words, sample size did not affect the stability of the equating results. However, there is some indication that they were slightly smaller for the longer tests in Studies 3 and 4 than in the shorter tests in Studies 1 and 2. Longer tests will normally improve the precision of the estimated equating results because they are likely to be more reliable and the correlation between the anchor items and the total test is likely to be higher.

Among the equating methods the CARC is the most stable method (smaller spread) across all the four studies and all sample sizes. However, it is also the least accurate after the IDEN method. Unlike the steady decline in the av.RMSD with increasing sample size that occurred for the NEAT methods and which was consistent in all four studies, the CARC showed some signs of following the same path but the decline was not consistent across all four studies. In study 4, its equating results became less accurate as the sample size increased from sample size 25 to 400. This result however, is not in contradiction with the notion that the CARC is specifically designed for small samples equating. The set of conditions in Study 4 (60 items, $a=1.0$) seems to favor this behavior on the part of CARC. Of the NEAT design methods the most accurate method is the SFRE. Next most accurate is the SCEE followed by TLIN, CLIN and last is the LLIN. The IDEN method is the worst among all of them both in terms of accuracy and stability and is independent of the sample size.

Overall, increasing sample size leads to more accurate equating results for NEAT methods in particular. The equating methods arranged in order of decreasing accuracy follow this sequence: SFRE, SCEE, TLIN, CLIN, LLIN, CARC and IDEN. The av.RMSD due to the NEAT methods share about the same variability but the results of the LLIN method tend to be slightly more stable.

Question 2: Interchangeability of Equating Results and av.RMSD

The chart in Figure 4.4 is used here to examine the degree of exchangeability of the equating results among the various methods. The plots in the panels in the first column (Study1) show, by the overlapping of the confidence bands that the NEAT methods are virtually exchangeable. However the LLIN becomes less exchangeable (because of it is relatively less accurate) with the other NEAT methods when the sample size is 25, which undoubtedly is very small. Moving to the second column (Study2), the methods are less exchangeable than they were under Study1. When the sample size is small (25) the NEAT methods are split into three clusters: LLIN, TLIN/CLIN and SFRE/SCEE. As sample size increases the LLIN method become more exchangeable with the other two linear methods giving rise to the “linear” cluster. For samples of size 50 or more, two clusters can be easily identified: SFRE/SCEE and LLIN/CLIN/LLIN. Although these differences can be seen from close inspection of the results, the actual magnitude of differences are small for most changes to sample size and across most methods (with only IDEN performing meaningfully worse than the other methods).

A slightly different type of clustering occurs in Study 3. All the NEAT methods with the exception of the LLIN method are now roughly equivalent. The exchangeability

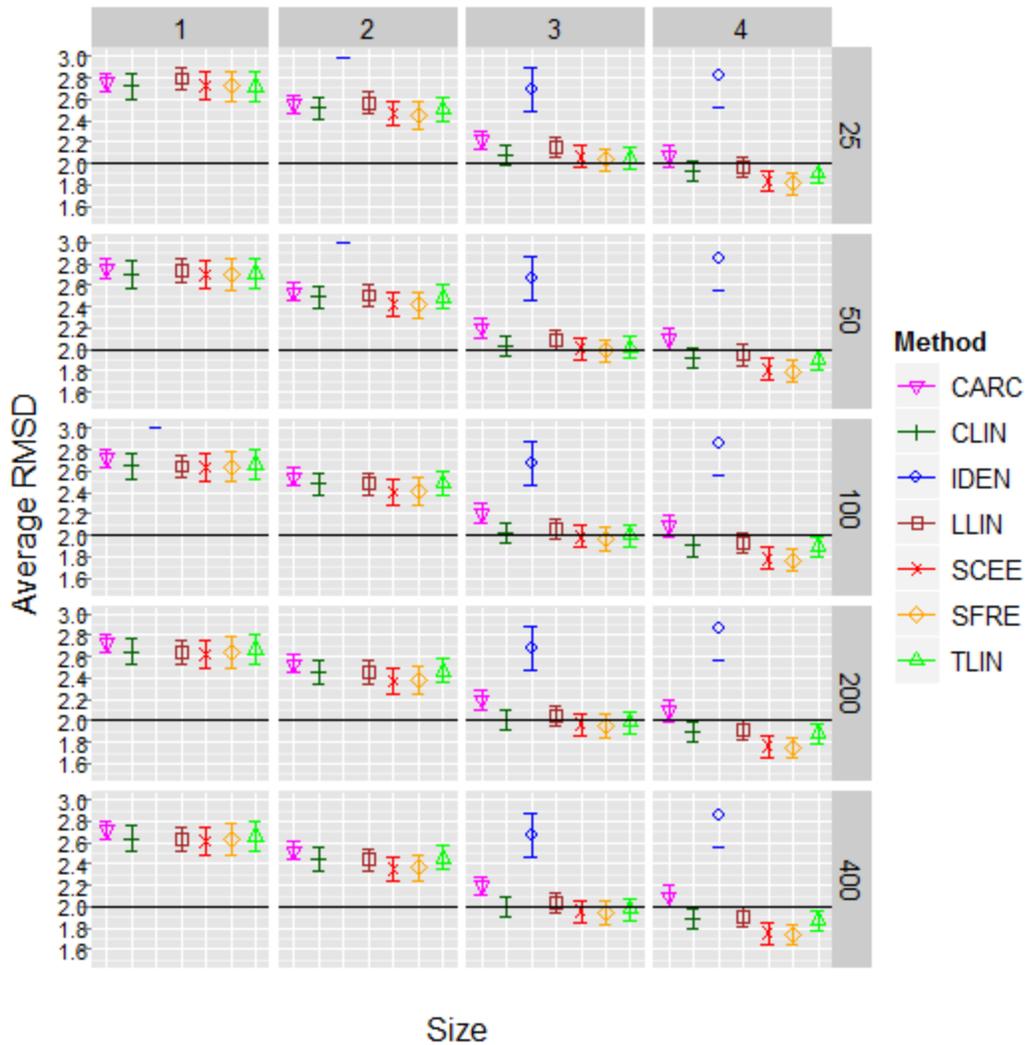


Figure 4.4: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various Sample Size by Studies Combinations

within the SFRE/SCEE or the TLIN/CLIN pair is now less solid. The SFRE is slightly but consistently more accurate than the SCEE. In turn the SCEE remains more exchangeable with the TLIN/CLIN pair. However, compared to the configurations in Studies 1 and 2, the two clusters (SFRE/SCEE and TLIN/CLIN) can be assumed to be

equivalent and considered as one single cluster. These results may also be considered to be close to those in Study 1 without the LLIN method.

The plots in Study4 are very similar to the plots in Study2. In the absence of the extremely small sample size, the formation of the SFRE/SCEE and TLIN/CLIN/ LLIN clusters can be clearly identified. Otherwise, if we take into account the very small sample size the LLIN method drops out of the list and only the two clusters (SFRE/SCEE and TLIN/CLIN) produce equating results that be considered exchangeable with one another. These results are identical to those in Study 2.

The formation of these clusters is directly related to the base form conditions and partially to sample size. Sample size is an only issue when the samples are close to or below what would be predicted by the central limit theorem. Under these extreme conditions the LLIN method becomes the least accuracy among the NEAT methods.

Other than sample size has little to no effect on the exchangeability of the equating results among the equating methods, another important conclusion is revealed from these findings: the base form average discrimination is a key factor that affects the exchangeability of the equating methods with one another with respect to the total equating error. The conclusions from the results described in the previous paragraphs can be summarized as: 1) the NEAT methods will all be virtually equivalent to one another if the base form average item discrimination is very low (0.6); 2) increasing the test length when the base discrimination is low makes the LLIN method less exchangeable with the other NEAT methods; 3) regardless of the test length increasing the base form average item discrimination to 1.0 splits the NEAT methods into two main clusters (“non- linear

smoothed” and the linear clusters); 4) increasing both the base forms test length and the average item discrimination (as in Study4) will accentuate the separation of the linear and “non-linear smoothed” clusters whereas decreasing both of them will draw the clusters together as if they were one; 5) the LLIN in general will fail to be part of the clusters whenever the sample size is 25 or 50.

The most important implication of these results is that the more exchangeable the equating results are among the NEAT equating methods based on the total equating error, chances are that the base form is not discriminating enough and changing the sample size will not change the outcomes. But increasing the base form discrimination to 1.0 will result in the formation of the expected clusters: the non-linear smoothing cluster and the linear cluster. One potential application of this result is that it might be used to perform diagnostics on the quality of the base forms and during the stages of test development or pre equating test forms.

As expected the equating results of the IDEN method are not exchangeable with any of the other methods whereas the CARC is only remotely exchangeable with the NEAT methods in Studies1 and 2 where the test forms contains only 30 items. It produces less accurate but more stable results than the NEAT methods, its equating results may be partially exchangeable with the other methods. Because this study is focused only on the aggregated accuracy statistics, it does not tell us where on the score scale it is likely to be most exchangeable with the other methods. This approach helps to compare the methods to one another in a general way under varying conditions and how these conditions affect the accuracy and consistency of their equating transformations.

Effect of Magnitude of Group Separation on Average BIAS

Question 1: Accuracy of Equating Methods and av.BIAS

This section analyses the results presented in Table 4.3 which relates to the av.BIAS of the various methods as function of the Standardized Theta Difference (STD) when pooled over all combinations of sample size, test length, α -ratio and Standardized Mean Difficulty Difference (SMD). Examination of the data from Table 4.3 shows that the av.BIAS of all methods (except the IDEN and CARC) is virtually near zero when there is no difference in the magnitude of the group separation. This means that if the alternate form sample and the base form sample are equally able, the equating results for all the NEAT equating methods will be virtually unbiased. This result is true across all four studies indicating that base form conditions may have little to no effect on the bias of the various methods when the groups are equally able. For all methods, except the CARC and IDEN, the average BIAS consistently increases in an almost exponential fashion as the STD between the forms increases, although, the STD condition is itself not uniform in its increases. That is, bias increases the most from $STD=.1$ to $STD=.25$, which is a .15 increase in group difference. There are smaller changes in bias as the STD increased from 0.0 to 0.05 or 0.10. In general, bias increases with STD. When $STD=0$ the range of the av. BIAS for the NEAT methods lies between 0.03 to 0.04 for Study1, -0.07 to -0.04 for Study2, -0.03 to -0.02 for Study3, and -0.03 to -0.01 for Study4. But when STD is 0.25 the range of values for the NEAT design methods lies between -0.46 to -0.09 and for Study1, -0.46 and -0.16 for Study 2, -0.33 and -0.08 for Study 3 and -0.08 and -0.29 for Study 4 respectively. These data indicate that large group separation has a

Table 4.3: Effect of STD on Average BIAS and SE of Equating Methods for Studies 1-4

<i>Study 1</i>	<i>STD</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
0		Mean	0.03	0.03	0.02	0.03	0.04	0.18	0.96
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.04
0.05		Mean	-0.03	-0.04	-0.07	-0.06	0.02	0.12	0.92
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.04
0.1		Mean	-0.15	-0.16	-0.22	-0.21	-0.07	0.01	0.88
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.04
0.25		Mean	-0.30	-0.31	-0.46	-0.46	-0.09	-0.12	0.82
		SE	0.03	0.03	0.03	0.03	0.04	0.05	0.20
<i>Study 2</i>	<i>STD</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
0		Mean	-0.05	-0.07	-0.05	-0.04	-0.07	0.23	1.30
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.05
0.05		Mean	-0.14	-0.16	-0.17	-0.16	-0.12	0.15	1.26
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.06
0.1		Mean	-0.12	-0.14	-0.18	-0.17	-0.06	0.16	1.28
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.05
0.25		Mean	-0.32	-0.34	-0.46	-0.45	-0.16	-0.01	1.18
		SE	0.03	0.03	0.02	0.02	0.03	0.06	0.23
<i>Study 3</i>	<i>STD</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
0		Mean	-0.02	-0.03	-0.02	-0.02	-0.02	0.65	0.98
		SE	0.02	0.02	0.01	0.01	0.02	0.04	0.19
0.05		Mean	-0.05	-0.06	-0.08	-0.07	-0.03	0.62	0.95
		SE	0.01	0.01	0.01	0.01	0.01	0.04	0.19
0.1		Mean	-0.10	-0.11	-0.15	-0.14	-0.05	0.57	0.93
		SE	0.02	0.02	0.02	0.02	0.02	0.04	0.19
0.25		Mean	-0.21	-0.21	-0.33	-0.32	-0.08	0.48	0.87
		SE	0.03	0.03	0.03	0.03	0.04	0.06	0.20
<i>Study 4</i>	<i>STD</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
0		Mean	-0.01	-0.03	-0.03	-0.01	-0.01	0.77	1.29
		SE	0.01	0.01	0.01	0.01	0.01	0.05	0.24
0.05		Mean	-0.04	-0.07	-0.08	-0.06	-0.02	0.74	1.27
		SE	0.01	0.01	0.01	0.01	0.01	0.06	0.23
0.1		Mean	-0.08	-0.11	-0.14	-0.12	-0.04	0.71	1.24
		SE	0.02	0.02	0.02	0.01	0.02	0.06	0.23
0.25		Mean	-0.19	-0.20	-0.29	-0.29	-0.08	0.61	1.17
		SE	0.03	0.03	0.03	0.03	0.03	0.06	0.23

very large impact on the av.BIAS of the various NEAT design methods. Even a relatively small group separation of .05, led to quite substantial increases in the av.BIAS and it would not be unreasonable to state that the differences in the av.BIAS corresponding to STD=0 and STD of 0.05, is significantly large. But in comparison with STD of .25, the

bias due to a STD of .05 is almost negligible. In fact, the average BIAS, when the STD is 0.05, is about less than a third the average BIAS when the STD is 0.25.

Figure 4.5 shows the distribution of the 95 percent CI of the av. BIAS as a function of STD for the various methods by studies combinations. The figure makes possible a visual comparison of the shifts in the estimate of the av. BIAS of the various methods as the STD changes. By comparing the panels along any one of four columns, it is obvious that the increase in the STD from 0.0 to 0.25 consistently drives the bias of the equating method deeper in the negative direction from the zero bias line. In other words as the alternate form group become more able relative to base form group the equating methods underestimate the equated scores. The one exception to this is the IDEN method which remains positively biased across all four studies.

Figure 4.6 shows the distribution of the 95 percent CI of the Average BIAS as a function of methods for the various STD by studies combinations. STD has a direct effect on the average bias of the equating methods, but at the same time the standard error (SE) of the estimates increases drastically as the magnitude of group separation increases. So not only does increasing group separation lead to equating results which are potentially very biased but these biased results can be very inconsistent. The immediate conclusion from these findings is that equating test forms for groups that are very disparate in ability can have very serious consequences in particular if the stakes are high because the equating results can both be inaccurate and unstable

In terms of the accuracy of the equating methods relative to one another, there is virtually no strict ordering when the groups are equally able because they are all almost

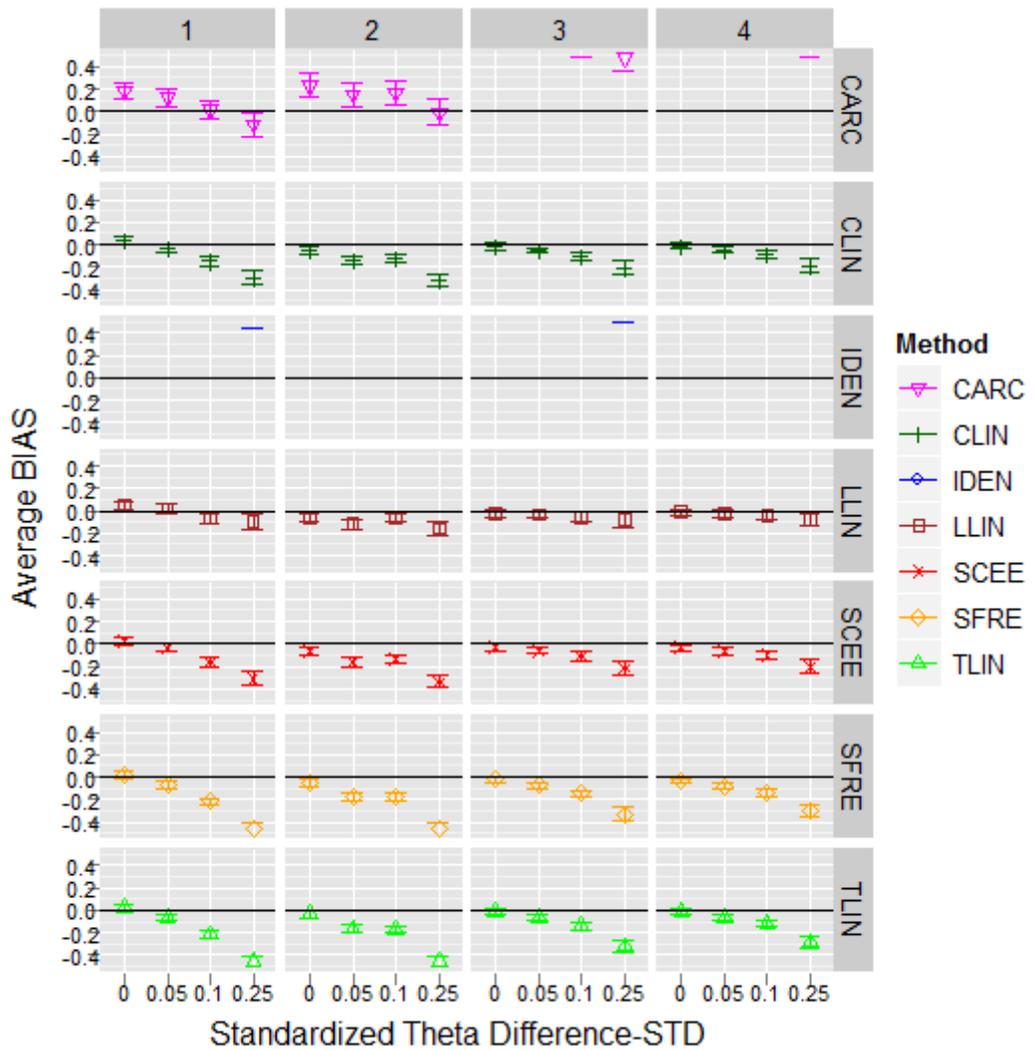


Figure 4.5: Distribution of the 95% CI of the Average BIAS as a Function of STD for the Various Methods by Studies Combinations

.equally biased. When the STD is not zero, the methods can be arranged in two different sequences arranged in order of least to most biased depending on the magnitude of group separation. The most obvious is the case when the STD is 0.1 or greater. The methods can be easily ordered as LLIN, CLIN, SCEE, TLIN, and SFRE though the CLIN/SCEE or TLIN/SFRE are virtually equally biased within each pair. The same ordering occurs

for the case when the $STD=0.05$ but the differences between the CLIN/SCEE and TLIN/SFRE pairs are not as pronounced as when the STD is 0.1 or larger. In other words, as these STD increases, the dissimilarity between these two pairs and with other methods also increases.

Examination of Figure 4.6 also shows that the equating methods, with the exception of the IDEN and CARC, become less biased as base form conditions changes from Study1 to Study 4. In this analysis the results of Studies 1 and 2 (30-item tests) are both less accurate and slightly more variable than the results of Studies 3 and 4 (60-item tests). This result is particularly true as long as the group separation is less than 0.1 STD . Scanning the panels across the rows of Figure 4.6 from left to right, reveals that doubling the test length tends to attenuate the effect of the large group separation on the av. BIAS. The attenuation effect is more pronounced when the magnitude of group separation become increasingly large. These results are primarily applicable to the SFRE, SCEE, TLIN and CLIN methods than to the LLIN, CARC and IDEN methods.

Note that the CARC is least biased when the STD is 0.1 in Study 1 and 0.25 in Study 2 and it is largely positively biased in both these studies when the STD is zero. Furthermore as the STD increases from zero to .25, there is a directional shift in the av.BIAS from relatively very positive to less positive (0.18 to -.12 in Study 1 or .23 to -.01 in Study 2).

Question 2: Interchangeability of Equating Results and av.BIAS

The findings of this section indicate that the NEAT methods may be classified into three groups: LLIN, CLIN/SCEE and TLIN/SFRE which merge or separate from one

another depending on the magnitude of separation (STD) of the alternate form group from the base forms group. Figure 4.6 shows that the equating methods, with the exception of the CARC and IDEN, are virtually exchangeable when the STD is zero. The three groups are joined as one and their equating results can be assumed as interchangeable among all five methods. In other words when the alternate form group and the base form group are equally able, the equating results of the various NEAT methods will be equally biased. A similar conclusion might apply to these methods when the STD is .05, except that the LLIN appears to be less exchangeable with the other NEAT methods, in particular the SFRE/TLIN pair. At this relatively small level of the STD (0.05), there is some indication that the two clusters SFRE/TLIN and CLIN/SCEE are still exchangeable to some degree based on the overlap of the 95% confidence bands.

However, when the STD is 0.1 or .25, the formation of the three clusters is clearly identified. There is practically no exchangeability among them as illustrated by the little overlap in their 95% CI in the bottom two rows of Figure 4.6. This result is true regardless of the base form conditions (i.e., across Studies 1-4) though one might argue that a longer test attenuates the effect of the large STD as indicated by the smaller average BIAS in Studies 3 and 4 compared to studies 1 and 2, when the STD is 0.1 or 0.25.

Furthermore, increasing the STD not only increases the bias of the equating results but it causes much larger increases in the bias of the equating results of the SFRE/TLIN pair than is produced relative to the CLIN/SCEE or LLIN methods. In other words, the STD accentuates the inexchangeability between the methods that tend to be most biased from those that tend to be more robust to variations in STD.

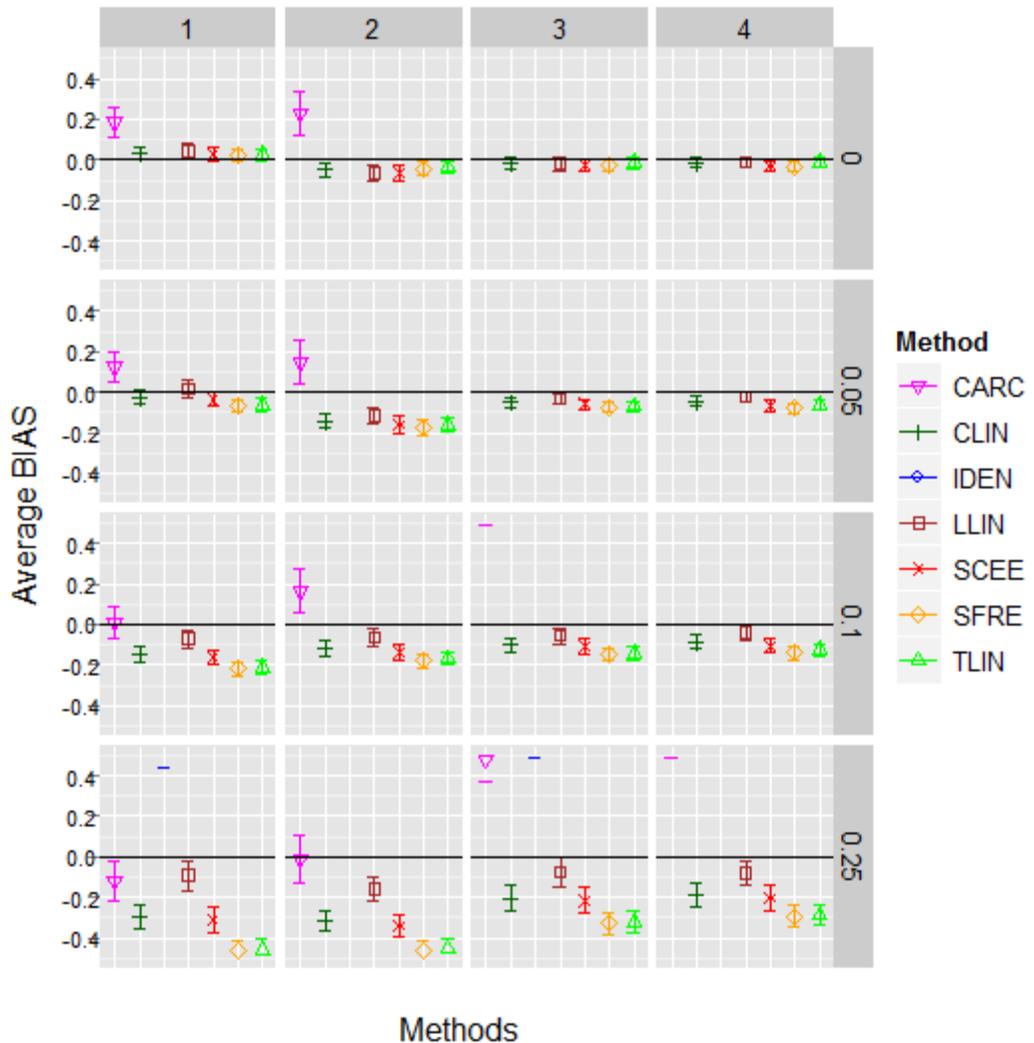


Figure 4.6: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various STD by Studies Combinations

To summarize, the effect of the STD on the average BIAS is beyond any ambiguity. As group separation increases the equating results become less accurate and less stable. At the same time the equating methods produce equating results that are less exchangeable with one another in terms of bias. The one element that remains unchanged in this analysis relates to the exchangeability within the SFRE /TLIN pair or the SCEE

/CLIN pair. The other important point is that all the NEAT methods are virtually exchangeable when the STD is zero and that as the STD reaches 0.1 or 0.25 the split into the three clusters, SFRE/TLIN, SCEE/CLIN, LLIN, become more well-defined.

Effect of Magnitude of Group Separation on Average RMSD

Question 1: Accuracy of Equating Methods and av. RMSD

This section analyses the results presented in Table 4.4 which relates to the average RMSD of the various methods as function of the Standardized Theta Difference (STD) when pooled over all combinations of sample size, test length, a-ratio and Standardized Mean Difference in Difficulty (SMD).

A comparison of the results from Table 4.4 shows within anyone of the four studies, the av.RMSD for the various methods (except of the CARC in Study4), is virtually constant across the four levels of STD. For example for the CLIN method, the STD ranges from 2.66 to 2.68 in Study 1, 2.47 to 2.50 in Study 2, 2.02 to 2.03 in Study 3 and 1.89 to 1.91 in Study 4. Relative to their associated standard errors (SE) the minor discrepancies in the av.RMSD across the four levels of the STD can be considered to be almost negligible. This finding suggests that on average the magnitude of the STD has practically no effect on the accuracy and stability of the overall equating results. In other words variations in the magnitude of the group separation do not have any significant impact on the (av.RMSD) overall equating error. These results are illustrated in Figure 4.7 which shows the distribution of the 95% CI of the av.RMSD as a function of STD for the various methods by Studies combinations. When the four studies are compared to one

Table 4.4: Effect of STD on Average RMSD and SE of Equating Methods for Studies1-4

<i>Study 1</i>	<i>STD</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
0		Mean	2.68	2.66	2.66	2.68	2.70	2.75	3.21
		SE	0.06	0.06	0.06	0.06	0.05	0.04	0.08
0.05		Mean	2.67	2.65	2.66	2.68	2.69	2.74	3.20
		SE	0.06	0.06	0.06	0.06	0.05	0.04	0.08
0.1		Mean	2.68	2.67	2.67	2.69	2.69	2.74	3.18
		SE	0.06	0.06	0.07	0.07	0.05	0.04	0.08
0.25		Mean	2.66	2.65	2.67	2.69	2.66	2.72	3.14
		SE	0.06	0.06	0.07	0.07	0.05	0.04	0.08
<i>Study 2</i>	<i>STD</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
0		Mean	2.50	2.41	2.41	2.49	2.52	2.56	3.30
		SE	0.05	0.05	0.05	0.05	0.04	0.04	0.12
0.05		Mean	2.47	2.39	2.39	2.47	2.48	2.54	3.28
		SE	0.05	0.05	0.06	0.05	0.05	0.04	0.12
0.1		Mean	2.46	2.40	2.40	2.47	2.48	2.53	3.27
		SE	0.05	0.05	0.05	0.05	0.05	0.04	0.12
0.25		Mean	2.47	2.40	2.42	2.49	2.48	2.52	3.21
		SE	0.05	0.06	0.06	0.05	0.05	0.04	0.11
<i>Study 3</i>	<i>STD</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
0		Mean	2.02	1.99	1.96	1.99	2.07	2.20	2.69
		SE	0.04	0.04	0.05	0.04	0.04	0.04	0.09
0.05		Mean	2.02	1.99	1.96	2.00	2.07	2.20	2.67
		SE	0.04	0.05	0.05	0.05	0.04	0.04	0.09
0.1		Mean	2.03	2.00	1.97	2.01	2.08	2.21	2.68
		SE	0.04	0.05	0.05	0.05	0.04	0.04	0.09
0.25		Mean	2.03	2.01	1.99	2.02	2.08	2.19	2.66
		SE	0.04	0.05	0.05	0.05	0.04	0.04	0.09
<i>Study 4</i>	<i>STD</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
0		Mean	1.91	1.79	1.77	1.89	1.94	2.11	2.87
		SE	0.04	0.05	0.04	0.04	0.04	0.04	0.14
0.05		Mean	1.91	1.79	1.77	1.90	1.94	2.10	2.87
		SE	0.04	0.05	0.04	0.04	0.04	0.04	0.13
0.1		Mean	1.91	1.79	1.77	1.89	1.94	2.09	2.86
		SE	0.04	0.05	0.04	0.04	0.04	0.04	0.13
0.25		Mean	1.89	1.78	1.77	1.88	1.91	2.04	2.79
		SE	0.05	0.05	0.05	0.05	0.05	0.04	0.13

another, there is clear indication from the lower av.RMSD values from Study4, that the test conditions of Study 4 are the most favorable for more accurate equating than Study 3 which in turn is better than Study2. The test conditions of Study 1 are without doubt the least suitable for accurate equating among the four Studies. Figure 4.7 clearly illustrates

the impact of test length on the av.RMSD in the presence of variations in the magnitude of group separation between the equated forms. The immediate conclusion from these findings is that doubling the test length significantly reduced the total equating error of all methods. At the same time the base form average item discrimination of 1.0 made the SFRE and SCEE methods more accurate than the other equating methods relative to one another. In particular, the “non-linear smoothing” methods became much more accurate than the “linear” methods.

In terms of the accuracy of the equating methods relative to one another, all the NEAT methods are virtually equal in accuracy under the worst base form condition in Study 1 (30 items, average item discrimination of 0.6). Even when the test length is doubled, the accuracy of the NEAT methods with the exception of the LLIN relative to one another is not altered drastically. If the SE is ignored the methods could be arranged as SFRE, SCEE, TLIN, CLIN, LLIN, CARC, IDEN but the differences are not as distinct between the SFRE, SCEE, TLIN, and CLIN when the base form discrimination is 0.6.

However, the base form discrimination of 1.0 accentuates the difference in the accuracy between the methods, making them more distinct from one another without altering the ordering described above. In other words, the ordering is not altered, the SFRE and SCEE remain virtually equally accurate just like the TLIN and CLIN methods. The main difference is that the SFRE and SCEE become much more accurate than the TLIN or CLIN methods as the base form discrimination increases to one.

Overall the findings suggest that the accuracy of the various methods relative to another is invariant to the STD but they do vary with the base form conditions. A longer

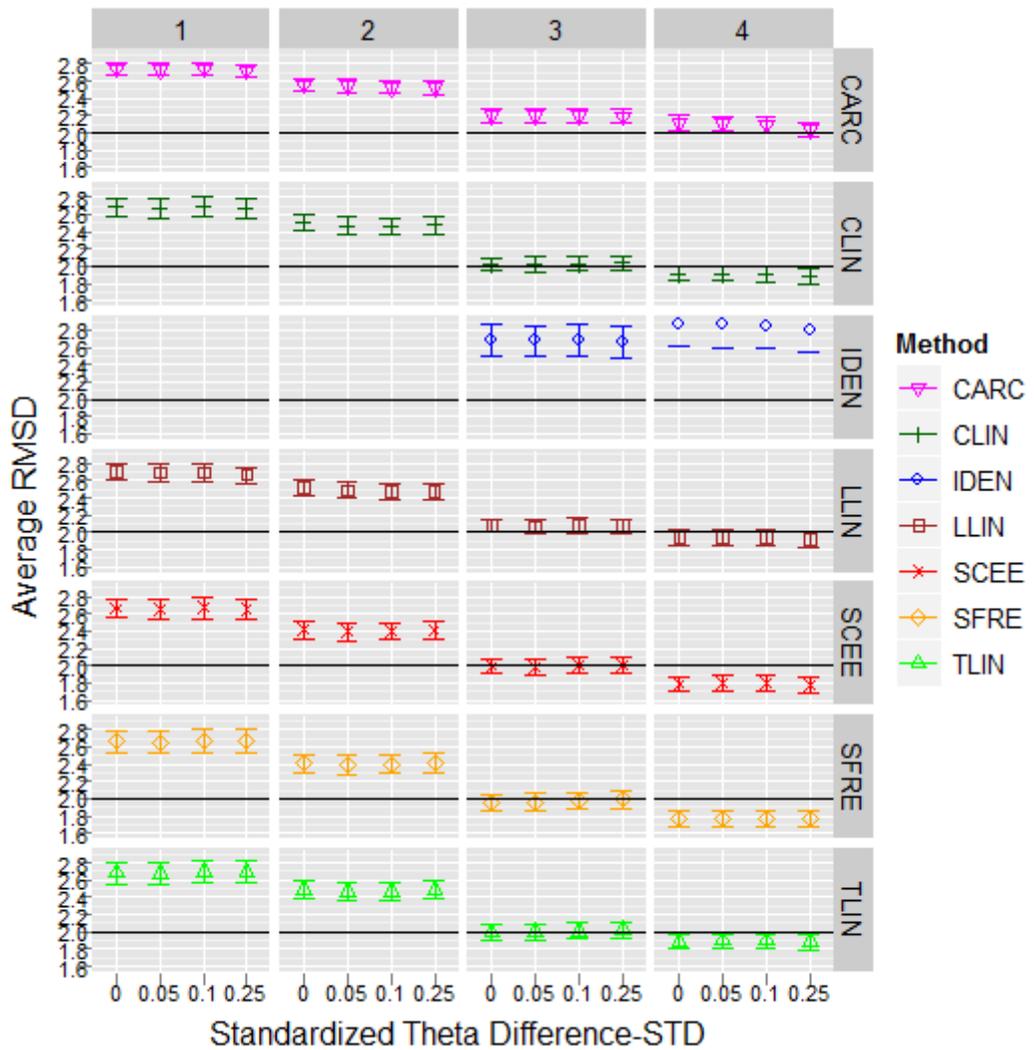


Figure 4.7: Distribution of the 95% CI of the Average RMSD as a Function of STD for the Various Methods by Studies Combinations

test with a more discriminating base will generally lower the overall equating error of all equating methods. The equating methods arranged in decreasing order of accuracy will generally follow this sequence: SFRE, SCEE, TLIN, CLIN, LLIN, CARC, IDEN. For example the average RMSD of the methods for the case when the STD is 0.05 in Study 4 is as follows: SFRE (1.77), SCEE (1.79), TLIN (1.90), CLIN (1.91), LLIN (1.94),

CRAC (2.10) and IDEN (2.87). This is essentially the general order the methods would be ranked based on the av.RMSD and it remains practically consistent across the four levels of STD.

Question 2: Interchangeability of Equating Results and av. RMSD

The findings of this section indicate that the NEAT methods may be thought of to consist of two groups: TLIN/ CLIN/ LLIN/ and SFRE/SCEE which merge or separate from one another depending on the test characteristics of the base forms. In the first group the LLIN method is slightly less accurate than the other two but its differences from the other two are not meaningfully large. Figure 4.8 displays the 95% CI of the av.RMSD from the data in Table 4.4 and presents the location of the various methods in relation to one another. It also shows the degree of exchangeability of the equating results among the various equating methods. The plots in the panels in the first column (Study1) show that the NEAT methods are virtually exchangeable with av.RMSD values ranging from 2.65 to 2.70. The equating methods can be thought of as a case when the two groups (CLIN/TLIN/LLIN and SFRE/SCEE) have come together as one, where the equating results of anyone equating method is exchangeable with the others with respect to the total equating error. In the second column (Study2), the methods are less exchangeable than they are under Study1. The SFRE and SCEE cluster together whereas the “linear” methods (CLIN/TLIN/LLIN) form another cluster of exchangeable methods. If we ignore the overlap in the 95% CI of their av.RMSD, these two clusters can be considered distinct and their equating results are relatively not exchangeable with one another. The plots in columns 3 (Study 3) have practically the same patterns as in the first column(Study 1)

except that the LLIN method is moderately detached from the other four NEAT methods. In Study 4, the patterns of exchangeability of the NEAT methods follow practically the same patterns as in Study 2.

Comparing the plots in Study 1 to Study 3 (30 items versus 60 items but same average base form average item discrimination of 0.6) or Study 2 to 4 (30 items versus 60 items but same average base form average item discrimination of 1.0) suggests that increasing test length essentially drives down the overall equating error of all methods, resulting in more accurate and more stable equating results

However, if we ignore the LLIN method it has practically no effect on the relative exchangeability of the other NEAT equating methods. More significant is the impact of the average item discrimination of the base forms on the exchangeability of the equated scores across equating methods. These results stem from the comparison of the panels in Studies 1 and 2 or Studies 3 and 4. The forms equated to the more discriminating base forms (1.0, Studies 2 or 4) stretch the NEAT methods further apart than the forms that are equated to the less discriminating (0.6, Studies 1 or 3) base forms. Its effect on the exchangeability of the equating results on the av.RMSD is definitely more noticeable than that of test length. Test length affects the accuracy but the average item-discrimination of the base forms affects the exchangeability and both of these two factors are invariant to changes in the STD.

Finally, the CARC, even as it happens to be generally the most stable among all the methods it remains consistently less accurate than the other methods, with the exception of the IDEN method. For its part the IDEN method is the simply the worst

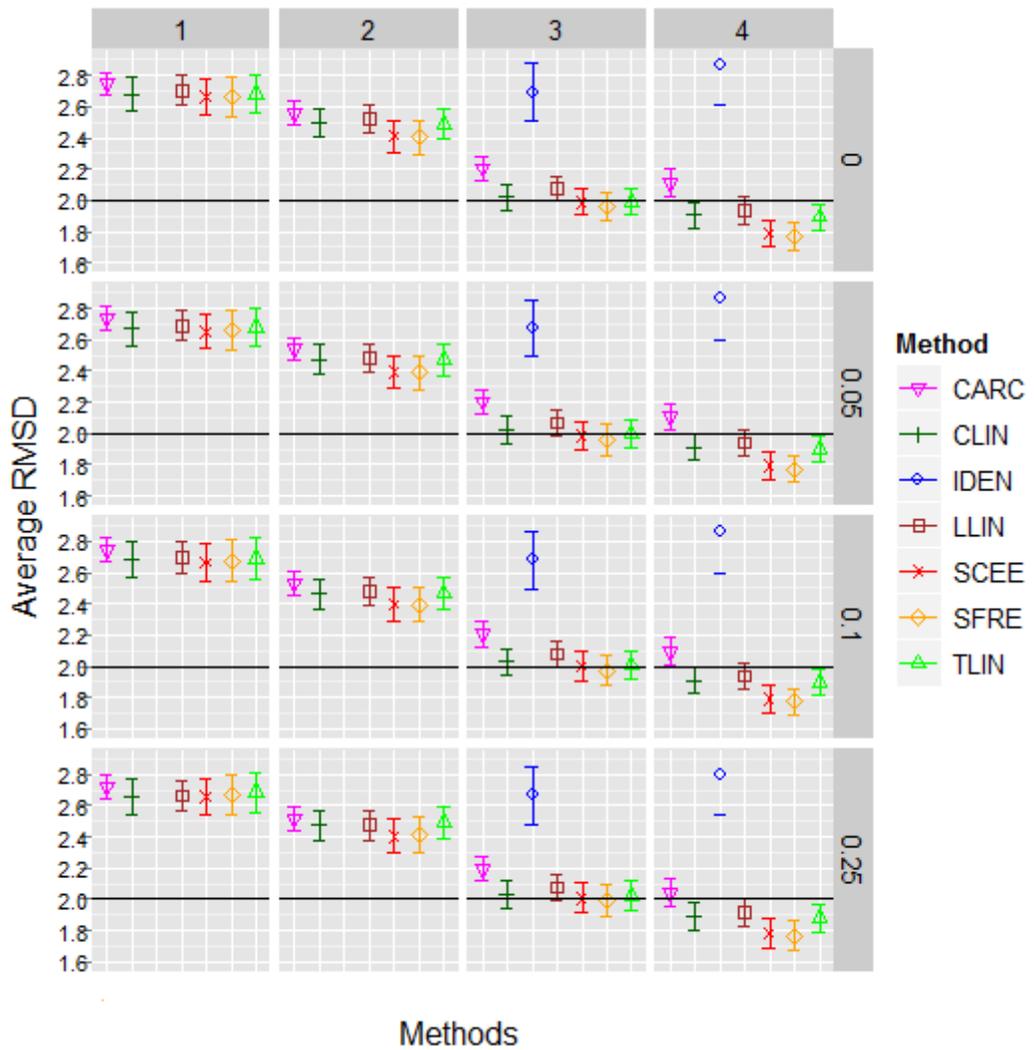


Figure 4.8: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various STD by Studies Combinations

method as indicated by its relatively high av.RMSD values and is not exchangeable with any of the other methods. The CARC on the other hand may be exchangeable with the LLIN method under only some unique conditions in Studies1 and 2 where the test forms contain 30 items.

Effect of Test Difficulty Differences (SMD) on Average BIAS

Question 1: Accuracy of Equating Methods and av.BIAS

This section analyses the results presented in Table 4.5 which relates to the av.BIAS of the various methods as function of the Standardized Theta Difference when pooled over all combinations of sample size, test length, α -ratio and Standardized Mean Difference. Examination of Table 4.5 reveals that the av.BIAS for the NEAT design methods is more or less constant across the various levels of SMD. For example the av.BIAS for the CLIN method for the 5 levels of SMD (-.75, -.5, -.1, 0, 0.25) are -.09, -.12, -.10, -.13, -.10 respectively. There is no consistent pattern that relates the av.BIAS to the various levels of SMD. Figure 4.9 illustrates the 95% CI of the av.BIAS as function of the SMD for all methods across the four studies. The av.BIAS of the NEAT design methods bounces up and down by very small amounts from one level of the SMD to the next. In fact, the differences between successive levels of the SMD to the next are not always large enough in relation to the standard error to really matter.

However, much stronger claims about the invariance of the av.BIAS with changes in the SMD can be made when the differences in form difficulty are less than -.50. This claim is more applicable to Studies 2 and 4 where the base form discrimination is 1.0 than to Studies 1 and 3 where the base form discrimination is 0.6. Overall there is enough evidence to suggest that the av.BIAS of the NEAT methods is independent of the SMD between the alternate forms and the base form, especially if the SMD is not exceedingly large or when the base form average item discrimination is one.

Table 4.5: Effect of SMD on Average BIAS and SE of Equating Methods for Studies1-4

<i>Study 1</i>	<i>SMD</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	-0.75	Mean	-0.09	-0.11	-0.17	-0.16	-0.01	0.43	3.22
		SE	0.02	0.02	0.03	0.03	0.02	0.05	0.12
	-0.50	Mean	-0.12	-0.13	-0.19	-0.18	-0.04	0.25	2.14
		SE	0.03	0.03	0.04	0.04	0.03	0.05	0.09
	-0.10	Mean	-0.10	-0.11	-0.18	-0.18	0.00	-0.01	0.40
		SE	0.03	0.03	0.03	0.03	0.03	0.03	0.04
	0	Mean	-0.13	-0.14	-0.20	-0.20	-0.06	-0.12	-0.07
		SE	0.03	0.03	0.03	0.03	0.03	0.03	0.03
	0.25	Mean	-0.10	-0.11	-0.17	-0.17	-0.01	-0.29	-1.19
		SE	0.03	0.03	0.03	0.03	0.03	0.03	0.03
<i>Study 2</i>	<i>SMD</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	-0.75	Mean	-0.18	-0.22	-0.26	-0.23	-0.13	0.68	4.13
		SE	0.03	0.03	0.03	0.03	0.03	0.06	0.11
	-0.50	Mean	-0.17	-0.19	-0.23	-0.21	-0.12	0.46	2.78
		SE	0.03	0.03	0.03	0.03	0.02	0.05	0.08
	-0.10	Mean	-0.14	-0.15	-0.19	-0.19	-0.08	0.03	0.64
		SE	0.02	0.02	0.03	0.03	0.02	0.03	0.03
	0	Mean	-0.15	-0.16	-0.20	-0.20	-0.09	-0.11	0.01
		SE	0.02	0.03	0.03	0.03	0.02	0.02	0.02
	0.25	Mean	-0.15	-0.16	-0.20	-0.20	-0.09	-0.40	-1.28
		SE	0.03	0.03	0.03	0.03	0.03	0.03	0.02
<i>Study 3</i>	<i>SMD</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	-0.75	Mean	-0.09	-0.11	-0.15	-0.13	-0.05	0.96	3.23
		SE	0.03	0.03	0.03	0.03	0.03	0.05	0.12
	-0.50	Mean	-0.08	-0.09	-0.13	-0.12	-0.02	0.80	2.18
		SE	0.03	0.03	0.03	0.03	0.03	0.05	0.09
	-0.10	Mean	-0.09	-0.09	-0.13	-0.13	-0.03	0.49	0.40
		SE	0.02	0.02	0.03	0.03	0.02	0.03	0.03
	0	Mean	-0.12	-0.13	-0.17	-0.16	-0.08	0.39	-0.03
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	0.25	Mean	-0.09	-0.10	-0.13	-0.13	-0.05	0.25	-1.12
		SE	0.02	0.02	0.02	0.02	0.03	0.02	0.03
<i>Study 4</i>	<i>SMD</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	-0.75	Mean	-0.07	-0.12	-0.15	-0.11	-0.03	1.35	4.12
		SE	0.02	0.02	0.02	0.02	0.02	0.06	0.10
	-0.50	Mean	-0.11	-0.13	-0.16	-0.15	-0.06	1.03	2.79
		SE	0.02	0.02	0.02	0.02	0.02	0.04	0.08
	-0.10	Mean	-0.09	-0.10	-0.13	-0.13	-0.05	0.55	0.58
		SE	0.02	0.03	0.03	0.03	0.02	0.03	0.03
	0	Mean	-0.08	-0.09	-0.12	-0.12	-0.04	0.44	0.04
		SE	0.03	0.03	0.03	0.03	0.03	0.03	0.02
	0.25	Mean	-0.06	-0.07	-0.11	-0.10	-0.02	0.17	-1.33
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.03

For the non-NEAT methods, CARC and IDEN, an inspection of the two right most columns of Table 4.5 reveals that these two methods are generally excessively biased except for one or two levels of the SMD. The CARC is least biased when the SMD is $-.10$ and the test is the shorter version of 30 items as in Studies 1 and 2. It is also less biased than the NEAT methods in the same studies when there is no difference in difficulty ($SMD=0$) between the equated forms. Note that of all the methods the IDEN and CARC showed some form of negative linear relationship with the SMD. For example with the CARC in Study 1, as the SMD increased across the five levels of SMD from $-.75$ to 0.25 its average BIAS changed from $.43$ to $.25$, $-.01$, $-.13$ and $-.29$. This trend is consistent for both methods across the four studies. By contrast the IDEN is consistently least biased when the SMD is zero regardless of the base form conditions. The actual data from Table 4.5 reads as: -0.07 for Study1, 0.01 for Study 2, -0.03 for Study 3, 0.04 for Study4. As indicated in the last paragraph, like the CARC, there is a swing in the direction of its bias, but unlike the CARC, it is most accurate when the $SMD = 0$ whereas the CARC is most accurate when the SMD is $-.10$.

The IDEN method is least biased among all the equating methods when the equated forms are equally difficult. It is also the least variable under the same conditions. In other words, as long as the equated forms are of equal difficulty ($SMD=0$) it produces the least biased and most stable equating results among all the methods. The last column in Table 4.5 clearly shows these results. Under all other SMD conditions the IDEN method is the most biased among all the methods. The results are clearly illustrated in the third row from the top of Figure 4.9 or the second row from the bottom of Figure 4.10.

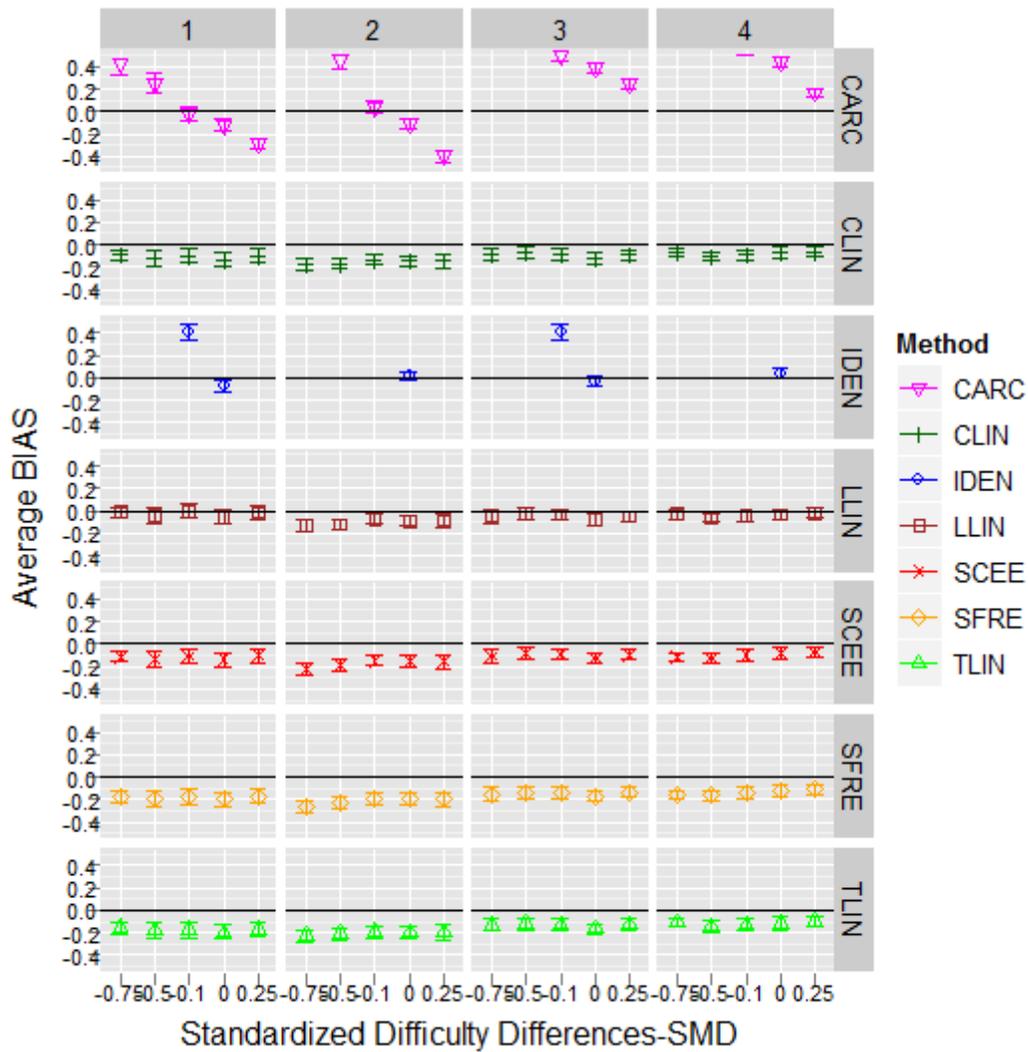


Figure 4.9: Distribution of the 95% CI of the Average BIAS as a Function of SMD for the Various Methods by Studies Combinations

This finding is simply a confirmation of a well established fact that the IDEN is extremely biased when the equated forms differ in difficulty but is essentially unbiased when they are equally difficult. Beyond this confirmation, these findings also indicate that all the other methods will be biased to some extent when the forms are equally difficult but the IDEN will not. For example in Study 2, the av.BIAS of the various

methods other than the IDEN when the SMD is zero ranges from -0.09 to -0.20 whereas the av.BIAS of the IDEN method is 0.01. This trend is virtually consistent across the four studies when the SMD is zero. As much as this might seem an odd finding, it is important to recall that the av.BIAS is aggregated over the range of all combinations of conditions when the SMD zero. In other words, the range of conditions includes extreme cases of group separation, highly unreliable alternate test forms and so forth. The most important point here is that the IDEN method is very robust to variations in test and sampling conditions as long as the alternate forms and the base forms are strictly parallel (i.e., SMD=0).

In terms of the performance of the various methods relative to one another, Figure 4.10 clearly shows that the LLIN method is in general the least biased among all the methods both within and across all four studies, regardless of the levels of SMD. The only exception to this rule is the IDEN method when the SMD is zero. Based on the plots in Figure 4.10, it seems that there is a consistent pattern which indicates that the LLIN method is generally the most accurate among them. It is followed in order of decreasing accuracy by the CLIN, SCEE, TLIN and SFRE. The IDEN is generally more biased except when the SMD is zero.

For the sake of completion, the effects of the base form characteristics in the presence of the SMD on the av.BIAS of the equating methods are quite complex. The effect of test length is not quite evident from the data in Table 4.5. It appears that test length interact with the equating methods. It affects some equating methods more than others. In comparing the data from Study 1 to Study3, the two most biased of the NEAT

methods (SFRE and TLIN) appear to be the greatest beneficiaries of the increase in test length whereas the other less biased methods show very little signs of any significant improvement. However, in Studies 2 and 4 all the NEAT methods improve with increasing test length.

Similarly, the effect of the base form average item discrimination is not consistent across Studies 1 and 2 or Studies 3 and 4. There is a substantial increase in the av.BIAS when the data in Study 2 is compared to Study 1 but the reverse occurs from Study 4 to Study 3. In other words, the effect of the base form discrimination has contradictory effects on the av.BIAS of the equating methods. In addition, the base form conditions interact with the SMD such that the av.BIAS of the NEAT methods in Studies 1 and 2 increases slightly (0 to 0.03 pts) for low levels of SMD and becomes worse (.05 or more points) with extreme values of the SMD. In Studies 3 and 4 they once more vary with the SMD but this time the av.BIAS decreases with low values and remains fairly the same for extreme SMD values.

The more likely reasonable conclusion is that the test length and the base form discrimination affect the av.BIAS in ways that depend on the levels of the SMD and the nature of the equating methods.

Question 2: Interchangeability of Equating Results and av.BIAS

In terms of exchangeability, the NEAT methods may be classified into three very distinct clusters: SFRE/TLIN, CLIN/SCEE, and the LLIN. The LLIN is virtually not exchangeable with any of the other NEAT methods though it is the most accurate among them. There is strong indication that, other than in the case when the SMD is extremely

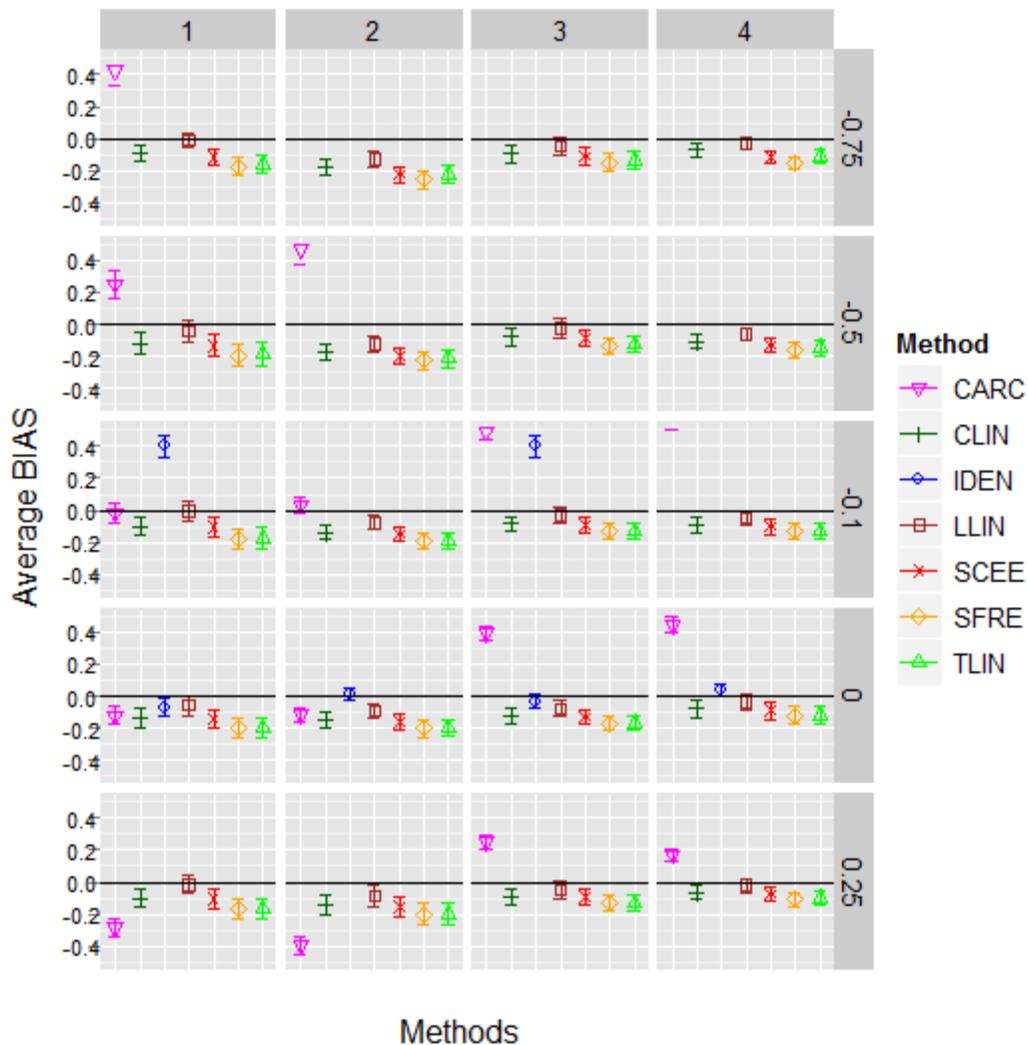


Figure 4.10: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various SMD by Studies Combinations

large (-.75), the demarcations between the three clusters are very clearly identified. These demarcations are present in the case of the extremely large SMD (-.75) provided the test consists of 30 items. If the test length is doubled and the SMD is -.75, the clusters as described above fail to form. Under these conditions, the equating results of the SCEE and TLIN methods are more exchangeable with each other than they are with their alter

ego. These results can be seen by comparing the top row of Figure 4.10 with the other rows.

Furthermore, the IDEN and CARC methods are generally not exchangeable with the other methods because they are simply exceedingly biased. They may be exchangeable with the LLIN method under very special circumstances when the base form is very short (30 items) and the SMD is very close to zero, otherwise their exchangeability can be ignored

Effect of Test Difficulty Differences on Average RMSD

Question 1: Accuracy of Equating Methods and av.RMSD

Analysis of Table 4.6 shows that the most accurate equating results occur when the SMD is small. From these results small may be defined as 0, -0.1 or 0.25. SMD values of -.5 and -.75 may be considered as large or extremely large. The plots in Figure 4.11 show the distribution of the 95% CI of the av.RMSD as the SMD increases from -.75 to .25 for all seven equating methods across Studies 1-4. It is obvious that there is quite a strong curvilinear relationship between the equating methods (except the SFRE and SCEE) and the SMD.

In general, except for the SFRE and SCEE methods, when the SMD is -.75 the total equating error is relatively very large and it decreases as the equated forms become more similar in difficulty and increases again as the SMD increases to 0.25. This relationship is true for all the equating methods across all four studies but more so for studies 2 and 4 than studies 1 and 3, i.e., both a longer test and a more discriminating

base form has the potential to cause much larger drops in the total equating error. The least error occurs when the alternate forms and the base forms are of equal difficulty. However, a difference of - 0.1 SMD does not produce significantly much larger equating errors and can be assumed to be close enough to no difference in difficulty. Even a difference of 0.25 SMD should not pose much problem with respect to the av.RMSD.

Among the equating methods, the IDEN method may produce relatively larger equating errors when the SMD is not very close to zero. If the base form conditions are very poor as in Study1 (30 items, $a=0.6$), and the SMD is -0.1, it may produce results that appear to be as accurate as the other methods. The same situation occurs when the SMD=0 and the base form conditions are poor. The problem when this happens is that these results may be interpreted as accurate and reliable when in fact the base form itself is the problem. Such a problem may easily go unnoticed because all the equating methods are producing equivalent equating results.

As the SMD increases to -0.5, the “linear” methods (LLIN, CLIN, TLIN) become much worse relative to when there is little to no difference in form difficulty. This is true particularly true for Studies 2, 3 and 4 whereas the same thing happens for all four studies when the SMD is -.75. By contrast the “non-linear smoothing” methods are more robust than the linear methods to variations in the SMD as illustrated in Figure 4.11. Actually when the SMD is extreme or large in Studies 1 or 3, their equating error is no different than when the equated forms are equal in difficulty. In studies 2 and 4, their equating error when the SMD is extremely large (-.75) is substantially higher by more than one standard error compared to the case when the SMD is equal to -.50.

Table 4.6: Effect of SMD on Average RMSD and SE of Equating Methods for Studies1-4

<i>Study 1</i>	<i>SMD</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	-0.75	Mean	2.72	2.65	2.65	2.73	2.74	2.85	4.19
		SE	0.05	0.06	0.07	0.06	0.05	0.03	0.06
	-0.50	Mean	2.65	2.62	2.63	2.66	2.66	2.73	3.41
		SE	0.06	0.06	0.07	0.07	0.05	0.04	0.03
	-0.10	Mean	2.67	2.67	2.68	2.68	2.68	2.68	2.69
		SE	0.07	0.07	0.08	0.07	0.06	0.05	0.04
	0	Mean	2.65	2.66	2.67	2.66	2.66	2.67	2.66
		SE	0.07	0.07	0.08	0.08	0.06	0.05	0.05
	0.25	Mean	2.68	2.69	2.70	2.69	2.69	2.76	2.96
		SE	0.07	0.07	0.08	0.08	0.06	0.04	0.03
<i>Study 2</i>	<i>SMD</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	-0.75	Mean	2.69	2.43	2.44	2.69	2.69	2.73	4.90
		SE	0.04	0.06	0.06	0.04	0.03	0.02	0.07
	-0.50	Mean	2.49	2.36	2.37	2.50	2.51	2.57	3.74
		SE	0.05	0.06	0.07	0.05	0.05	0.03	0.03
	-0.10	Mean	2.39	2.38	2.38	2.40	2.41	2.41	2.49
		SE	0.06	0.06	0.06	0.06	0.05	0.05	0.04
	0	Mean	2.38	2.39	2.39	2.38	2.40	2.43	2.41
		SE	0.06	0.06	0.06	0.06	0.06	0.05	0.05
	0.25	Mean	2.42	2.44	2.44	2.43	2.44	2.54	2.77
		SE	0.06	0.06	0.06	0.06	0.06	0.04	0.04
<i>Study 3</i>	<i>SMD</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	-0.75	Mean	2.10	2.00	1.97	2.08	2.15	2.44	3.89
		SE	0.04	0.05	0.05	0.04	0.04	0.03	0.07
	-0.50	Mean	2.03	1.98	1.95	2.01	2.08	2.28	3.04
		SE	0.05	0.05	0.06	0.05	0.04	0.03	0.03
	-0.10	Mean	1.99	2.00	1.97	1.97	2.04	2.12	2.08
		SE	0.05	0.05	0.06	0.06	0.05	0.04	0.05
	0	Mean	1.98	1.99	1.96	1.96	2.03	2.07	2.02
		SE	0.05	0.05	0.06	0.06	0.05	0.05	0.05
	0.25	Mean	2.02	2.02	2.00	1.99	2.06	2.09	2.36
		SE	0.05	0.05	0.05	0.05	0.05	0.05	0.03
<i>Study 4</i>	<i>SMD</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	-0.75	Mean	2.19	1.82	1.80	2.18	2.21	2.50	4.70
		SE	0.02	0.05	0.05	0.02	0.02	0.02	0.07
	-0.50	Mean	1.96	1.76	1.74	1.94	1.98	2.20	3.46
		SE	0.04	0.05	0.05	0.04	0.04	0.03	0.04
	-0.10	Mean	1.78	1.77	1.75	1.77	1.81	1.93	1.94
		SE	0.05	0.05	0.05	0.05	0.05	0.05	0.04
	0	Mean	1.77	1.77	1.76	1.76	1.80	1.89	1.82
		SE	0.05	0.05	0.05	0.05	0.06	0.05	0.05
	0.25	Mean	1.82	1.82	1.80	1.81	1.85	1.91	2.32
		SE	0.05	0.05	0.05	0.05	0.05	0.05	0.03

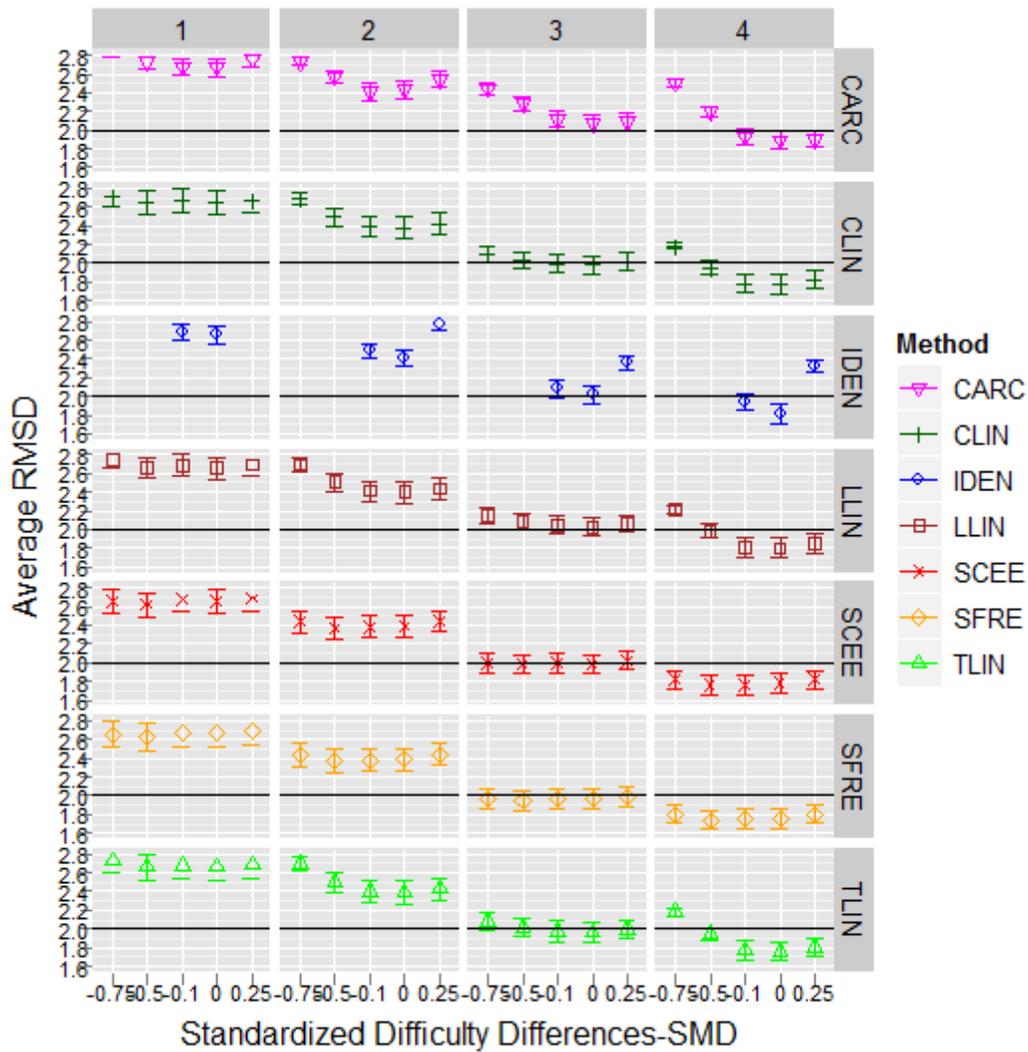


Figure 4.11: Distribution of the 95% CI of the Average RMSD as a Function of SMD for the Various Methods by Studies Combinations

Clearly, the base form conditions matter most when the SMD values are large or extremely large. A low base form average item discrimination (0.6) as in studies 1 and 3 led the “non –linear smoothing” methods to produce the same total equating error across all levels of the SMD. By contrast a higher base form average item discrimination (1.0)

as in Studies 2 and 4 caused these methods to produce the same total equating error across all levels of the SMD except when the SMD was $-.75$.

The “linear” methods on the other hand, are more susceptible to the effect of large differences in form difficulty. A low base form average item discrimination (0.6) as in studies 1 and 3 led the “linear” methods to produce the same total equating error across all levels except when the SMD is $-.75$. By contrast a higher base form average item discrimination (1.0) as in Studies 2 and 4 caused the “linear” methods to produce the same total equating error across all levels of the SMD except when the SMD was $-.50$ or $-.75$.

In those instances when the SMD is $-.75$ or $-.50$ the discrepancy in the total equating error of the “linear” methods is so large that these methods cannot make the adjustment to bring it back to the same level as when the SMD is small. Clearly the “non-linear smoothing” methods are more capable of making larger adjustments than the “linear” methods. The most affected by the large SMD are the linear methods.

In terms of the accuracy of the equating method relative to one another, there is no strict sequence they can be arranged into when the SMD is 0 or $-.10$ because all the equating methods produce about the same amount of equating error. But when the SMD is large, the arrangement is more definite and the sequence in order of decreasing accuracy can be assumed to be: SFRE, SCEE, TLIN, CLIN, LLIN, CARC, IDEN.

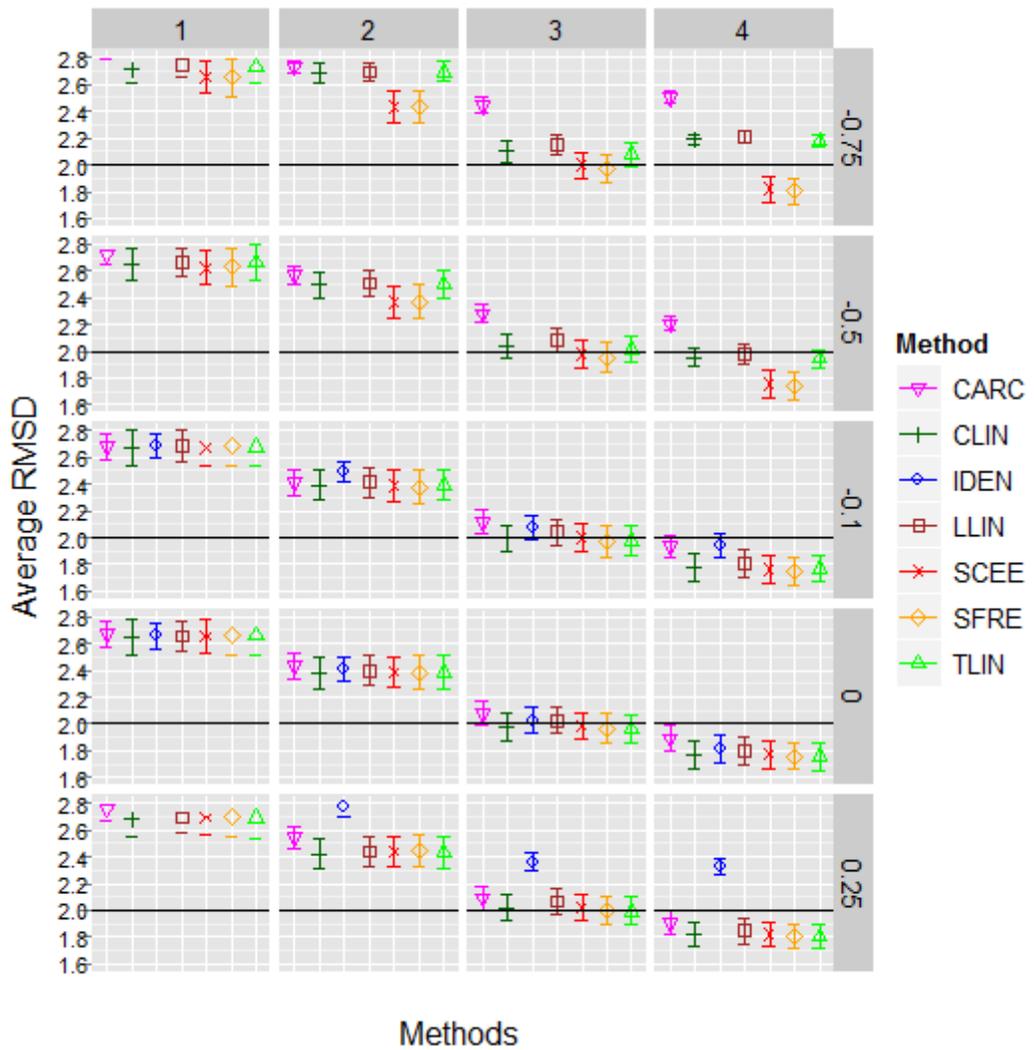


Figure 4.12: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various SMD by Studies Combinations

Question 2: Interchangeability of Equating Results and av.RMSD

Figure 4.12 displays the 95% CI of the av.RMSD from the data in Table 4.6 and presents the location of the various methods in relation to one another. All the methods are virtually exchangeable when the SMD is small (0, -.1, .25). The exception is the IDEN which is exchangeable with the other methods only when the SMD is zero. When

the SMD is large or extreme the NEAT methods are split into two clusters: linear and “smoothing”. The LLIN method in the linear cluster hangs with the CLIN and TLIN if the test is short as in studies 1 or 2. Otherwise when the test length is doubled it is less accurate and is not as close as the other two linear methods are to each other.

Overall, the exchangeability of the equating results among the equating methods is greatest when the SMD is zero and progressively diminishes as the differences in form difficulty increases. The one constant is that the "smoothing" and the TLIN/CLIN pairs are the two main clusters within which the equating methods are virtually exchangeable. Under less strict conditions the LLIN method conditions might also be considered exchangeable with the other linear methods.

By comparing the panels in columns 1 and 3 or columns 2 and 4 in Figure 4.12, it is clear that test length has the most effect on the exchangeability of the equating results when the SMD is -0.5 or -.75. The “non-linear smoothing” methods move further away from the other methods. The base form discrimination accentuates the separation even more as long as the SMD exceeds .25. If the SMD is small, test length or the base form discrimination has barely any impact on the exchangeability of the equating results for nearly all the methods except the IDEN method. As mentioned earlier the IDEN method is most accurate and exchangeable with the other methods when the test forms are equally difficult.

Effect of Discrimination Ratio on Average BIAS

Question 1: Accuracy of Equating Methods and av.BIAS

This section analyses the results presented in Table 4.7 which relates to the av.BIAS of the various methods as function of the discrimination ratio of the new form to the base form, when pooled over all combinations of sample size, test length, STD and SMD.

Table 4.7: Effect of a-ratio on Average BIAS and SE of Equating Methods for Studies1-4

<i>Study 1</i>	<i>a-ratio</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	0.3/0.6	Mean	-0.23	-0.25	-0.28	-0.28	-0.17	-0.17	0.42
		SE	0.03	0.03	0.03	0.03	0.03	0.03	0.11
	0.6/0.6	Mean	-0.07	-0.08	-0.15	-0.15	0.02	0.09	1.00
		SE	0.02	0.02	0.02	0.02	0.02	0.03	0.17
	0.9/0.6	Mean	-0.03	-0.03	-0.12	-0.11	0.07	0.23	1.28
		SE	0.02	0.02	0.02	0.02	0.02	0.04	0.20
<i>Study 2</i>	<i>a-ratio</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	0.5/1.0	Mean	-0.26	-0.29	-0.31	-0.29	-0.22	-0.11	0.84
		SE	0.02	0.02	0.03	0.03	0.02	0.03	0.15
	1.0/1.0	Mean	-0.12	-0.13	-0.18	-0.18	-0.06	0.19	1.37
		SE	0.02	0.02	0.02	0.02	0.02	0.04	0.21
	1.5/1.0	Mean	-0.09	-0.10	-0.16	-0.15	-0.03	0.31	1.56
		SE	0.01	0.01	0.02	0.02	0.02	0.06	0.23
<i>Study 3</i>	<i>a-ratio</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	0.3/0.6	Mean	-0.21	-0.23	-0.24	-0.24	-0.18	0.35	0.46
		SE	0.02	0.02	0.03	0.03	0.02	0.03	0.11
	0.6/0.6	Mean	-0.09	-0.09	-0.13	-0.13	-0.04	0.59	1.01
		SE	0.01	0.01	0.01	0.01	0.01	0.03	0.17
	0.9/0.6	Mean	0.01	0.01	-0.06	-0.04	0.08	0.80	1.32
		SE	0.01	0.01	0.01	0.01	0.02	0.04	0.20
<i>Study 4</i>	<i>a-ratio</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	0.5/1.0	Mean	-0.19	-0.23	-0.25	-0.22	-0.15	0.46	0.85
		SE	0.02	0.02	0.02	0.02	0.02	0.03	0.15
	1.0/1.0	Mean	-0.05	-0.06	-0.10	-0.09	0.00	0.77	1.34
		SE	0.01	0.01	0.01	0.01	0.01	0.05	0.21
	1.5/1.0	Mean	-0.01	-0.01	-0.06	-0.05	0.04	0.90	1.53
		SE	0.01	0.01	0.01	0.01	0.01	0.06	0.23

Examination of Table 4.7 reveals that within anyone of the four studies, the largest bias among the various methods occurs (except for the CARC and IDEN) when the a-ratio is 0.5, i.e., the average discrimination of items on the new form is half that on the old form. Inspection of the data in the first row of each of the four studies shows that the range of values for the case when the a-ratio is 0.5 is relatively very large compared to the data in the third or fifth rows. The third row shows the bias for the case when the a-ratio is 1.0 and the fifth row shows the data for the case when the a-ratio is 1.5.

The differences across the rows are strikingly large and suggest that the a-ratio can have serious consequences on the equating results in terms of the bias. Restated differently, equating alternate forms which are much less discriminating than the base form average item discrimination can cause severe distortions in the equating results. These differences due to item quality across the forms are much larger than the differences attributed to the use of different equating methods.

Based on the data in Table 4.7, the equating results will be severely underestimated if the ratio is 0.5 than if the ratio is 1.0 and are even worse compared to an a-ratio of 1.5. In all four studies the av.BIAS is smallest for all methods (except the CARC and IDEN) when the a-ratio is 1.5. Based on the SE, these differences in the av.BIAS from one level of the a-ratio to the next cannot be considered as trivial. In general when the a-ratio is 0.5 (i.e., discrimination of alternate forms are 0.3 or 0.5) all the equating methods, except the CARC, produce equating results that are biased by a factor that is at least almost double or more than double in some cases that would occur when the ratios are 1.0 or 1.5 (i.e., when the discrimination is 0.6 or 0.9 if the base form discrimination is

0.6 or 1.0 or 1.5 when the base form average discrimination is 1.0). In other words the bias incurred by all the methods in the NEAT design group is cut by at least half or more as the average item discrimination of the alternate forms changes from 0.3 to 0.6 or 0.9 in studies 1 and 3 and from 0.5 to 1.0 to 1.5 in studies 2 and 4.

A graphical representation of the 95% CI of the av.BIAS as a function of the a-ratio computed for the various methods based on Table 4.7 is illustrated in Figure 4.13. The plots in the figure indicate very clearly that the equating results of all the equating (except the CARC and the IDEN) methods tend to become less biased as the discrimination ratio of the alternate form to the base form increases from 0.5 to 1.5.

In terms of the relative bias of the various methods, the plots in Figure 4.13 clearly show that the NEAT methods are nearly parallel to another in general. However, the LLIN method is not always consistently the least biased methods as was the case in the findings in the previous sections. Only under the condition that the a-ratio is 0.5 or 1.0, does the LLIN method remain consistently the least biased among all the methods. Otherwise when the a-ratio is 1.5 and the test is doubled in length as in Studies 3 or 4, the CLIN and SCEE, (CE), methods are more accurate than the LLIN method.

Overall, from Figure 4.14 the NEAT methods arranged in order of increasing av.BIAS yields two sequences which depend on the a-ratio. The first sequence occurs when the a-ratio is 0.5 or 1 and can be arranged as LLIN, CLIN, SCEE, TLIN and SFRE.

The other sequence occurs when the a-ratio is 1.5 and can be arranged as CLIN, SCEE, LLIN, TLIN and SFRE. Note how the LLIN method has moved from being the least biased in the first arrangement to be in the middle in the second sequence. The

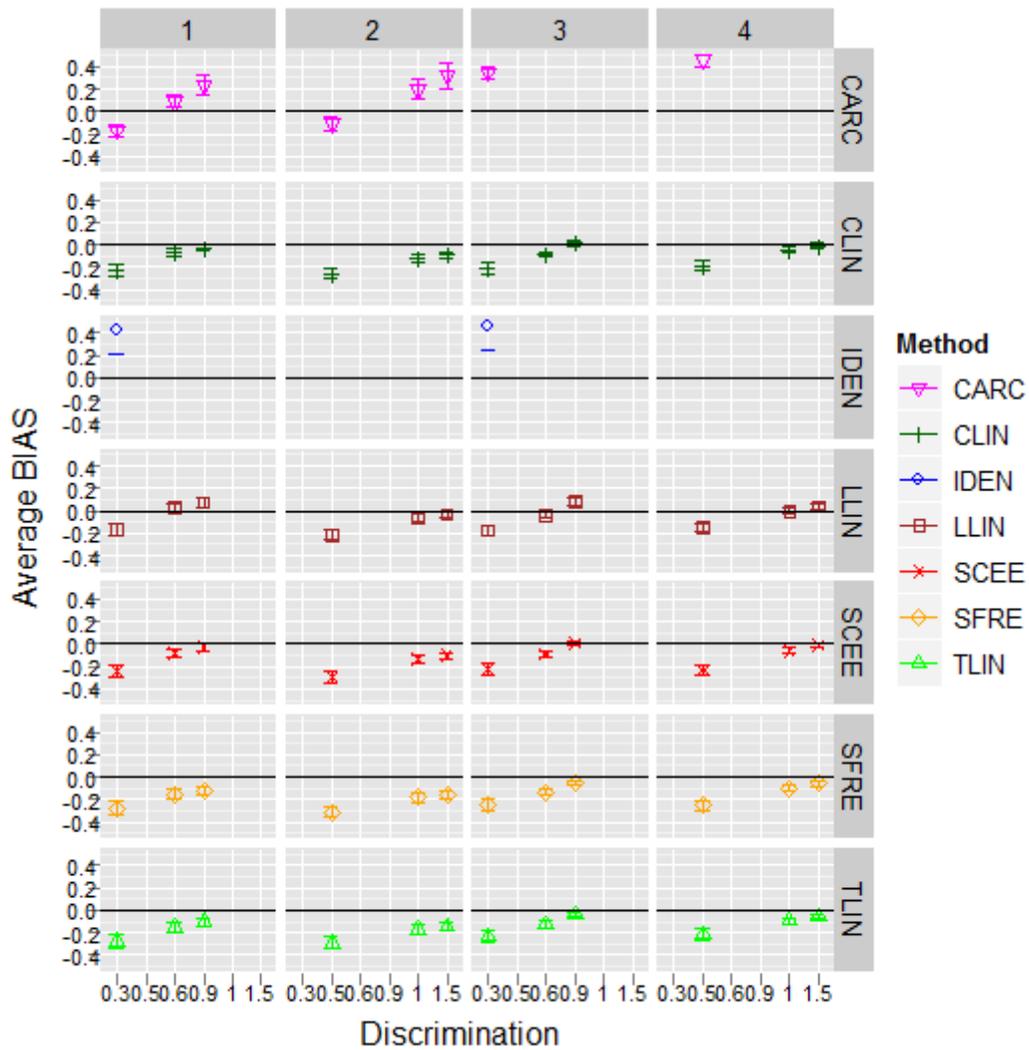


Figure 4.13: Distribution of the 95% CI of the Average BIAS as a Function of Discrimination (a-ratio) for the Various Methods by Studies Combinations

av.BIAS of the CARC and IDEN for their part method are exceeding large across all levels of the a-ratio. One exception will be the case of the CARC when the a-ratio is small (0.5) and the test length is short as in Studies 1 and 2. Under these circumstances it is the least biased method. Does this suggest that the CARC method perform better when

there is little test information? This seems to support similar findings about the CARC with respect to short test length and low base form discrimination in the earlier sections.

Furthermore from the data in Table 4.7 and comparing the plots in Figure 4.14 across all four studies it appears that the base form conditions have a direct effect on the magnitude of the av.BIAS of the equating transformations. Comparing the results of Study 3 to Study1 and repeating the same comparisons between Study 4 to Study 2 gives a sense of the effect of test length in the presence of a-ratio. Comparing the results of Study 2 to Study1 and repeating the same comparisons between Study 4 to Study3 gives a sense of the effect of the base form average item discrimination in the presence of a-ratio. There is definitely a larger gain in accuracy from Study 2 to Study4 (which represents a change in test length from 30 to 60 items but with the base form average item discrimination set at 1.0) for most of the NEAT methods at all levels of the SMD than from Study1 to Study3 (which represents a change in test length from 30 to 60 items but with the base form average item discrimination set at 0.6). On the other hand, the NEAT equating methods appeared to become more biased from Study1 to Study2 and less biased from Study 3 to Study 4, at least for some levels of the SMD. One possible explanation for these inconsistencies is that the base forms conditions interact with the a-ratio.

Question 2: Interchangeability of Equating Results and av.BIAS

In terms of exchangeability of the equating results the same clustering of the methods that was observed in the previous sections with respect to the av.BIAS is present here as can be identified from Table 4.7. The TLIN /SFRE and CLIN/SCEE pairings are

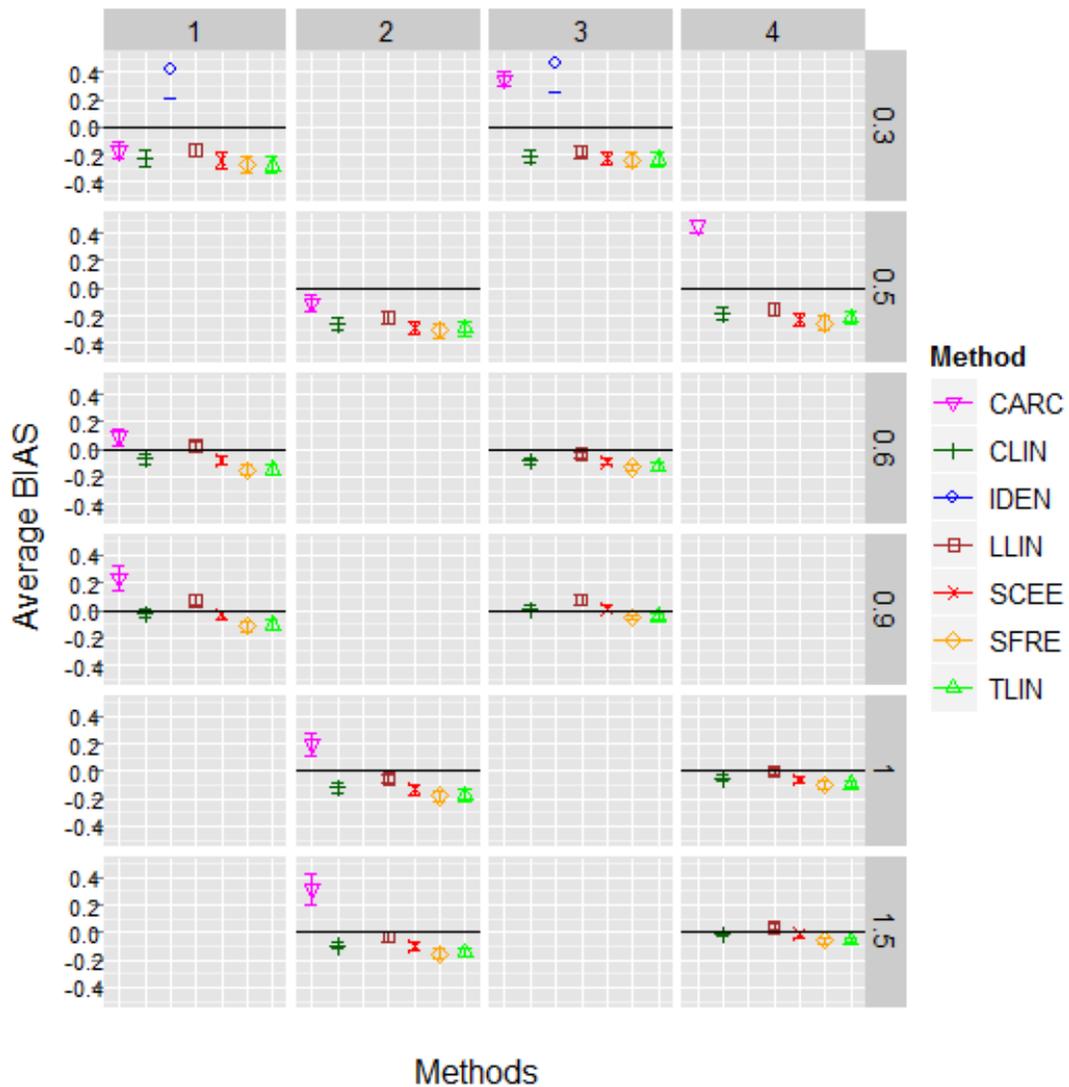


Figure 4.14: Distribution of the 95% CI of the Average BIAS as a Function of Methods or the Various Form Discrimination (a-ratio) by Studies Combinations

present under most conditions. However, when the a-ratio is 0.5, the distinction between these two clusters is not as pronounced as when the a-ratio is 1.0 or higher. Note that the base form conditions do not appear to affect the degree of exchangeability between the methods within the TLIN/SFRE and CLIN/SCEE clusters and these two clusters are

nearly almost not exchangeable with one another. For example in the case of the a-ratio in Study 4 the av.BIAS is -0.05 and -0.06 for the CLIN/SCEE pair, -.10 and -0.09 for the SFRE/TLIN pair with the SE solidly centered at 0.01. This pattern is consistent across all four studies, except when the a-ratio is 0.5.

Effect of Discrimination Ratio on Average RMSD

Question 1: Accuracy of Equating Methods and av.RMSD

This section analyses the results presented in Table 4.8 which relates to the av.RMSD of the various methods as function of the discrimination ratio, when pooled over all combinations of sample size, test length, Standardized Theta Difference and Standardized Mean Difference.

Examination of Table 4.8 reveals that within anyone of the four studies, the largest equating error, in terms of the av.RMSD values, among the various methods occurs (except for the CARC and IDEN) when the a-ratio=0.5 (i.e., the average discrimination of items on the new form is half that on the old form). The first row of each of the four studies in the table shows the av.RMSD values for each of the equating methods for the case when the a-ratio is 0.5. The third row shows the errors for the case when the a-ratio is 1.0 and the fifth row refers to the case when the a-ratio is 1.

As the a-ratio increases from 0.5 to 1.5 the av.RMSD for all methods, except the IDEN and CARC, consistently drops in a monotonic fashion within all four studies. For instance in Study1 when the a-ratio is 0.5 the av.RMSD due the CLIN method is 3.34; it is 2.46 when the a-ratio is 1.0 and 2.22 when the a-ratio is 1.5. Similar results occur for

Table 4.8: Effect of a-ratio on Average RMSD and SE of Equating Methods for Studies 1-4

<i>Study 1</i>	<i>a-ratio</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	0.3/0.6	Mean	3.34	3.34	3.43	3.42	3.25	3.16	3.31
		SE	0.01	0.01	0.01	0.01	0.01	0.01	0.02
	0.6/0.6	Mean	2.46	2.45	2.45	2.46	2.47	2.48	3.00
		SE	0.01	0.01	0.01	0.01	0.01	0.01	0.07
	0.9/0.6	Mean	2.22	2.18	2.11	2.16	2.34	2.56	3.23
		SE	0.01	0.01	0.01	0.01	0.01	0.02	0.09
<i>Study 2</i>	<i>a-ratio</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	0.5/1.0	Mean	3.02	3.02	3.06	3.05	3.01	2.94	3.34
		SE	0.01	0.01	0.01	0.01	0.01	0.01	0.05
	1.0/1.0	Mean	2.29	2.21	2.21	2.29	2.30	2.32	3.14
		SE	0.01	0.01	0.01	0.01	0.01	0.02	0.11
	1.5/1.0	Mean	2.12	1.97	1.94	2.10	2.16	2.35	3.30
		SE	0.02	0.01	0.01	0.02	0.02	0.02	0.13
<i>Study 3</i>	<i>a-ratio</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	0.3/0.6	Mean	2.54	2.54	2.56	2.55	2.56	2.57	2.77
		SE	0.01	0.01	0.01	0.01	0.01	0.01	0.03
	0.6/0.6	Mean	1.78	1.76	1.75	1.77	1.79	1.87	2.45
		SE	0.01	0.01	0.01	0.01	0.01	0.02	0.08
	0.9/0.6	Mean	1.76	1.69	1.60	1.69	1.87	2.16	2.81
		SE	0.01	0.01	0.01	0.01	0.01	0.03	0.10
<i>Study 4</i>	<i>a-ratio</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
	0.5/1.0	Mean	2.35	2.33	2.30	2.33	2.39	2.42	2.91
		SE	0.01	0.01	0.01	0.01	0.01	0.01	0.06
	1.0/1.0	Mean	1.71	1.59	1.59	1.71	1.71	1.85	2.72
		SE	0.02	0.01	0.01	0.02	0.02	0.03	0.13
	1.5/1.0	Mean	1.66	1.44	1.42	1.64	1.69	1.99	2.91
		SE	0.03	0.01	0.01	0.03	0.03	0.04	0.14

the other methods in all four studies. However, the av.RMSD values are relatively higher in Studies 1 and 2 where they are all higher than the 2 point mark. In studies 3 and 4 the av. RMSD values are lower than 2 points only when the a-ratio between the new form and the base form is 1.0 or higher, otherwise the av.RMSD values are higher than the 2 point mark when the a-ratio is 0.5. Taking into account the SE of the average RMSD, the differences across the different levels of a-ratio for any method cannot be considered as trivial. To support this finding a graphical representation of the 95% CI of the av.RMSD

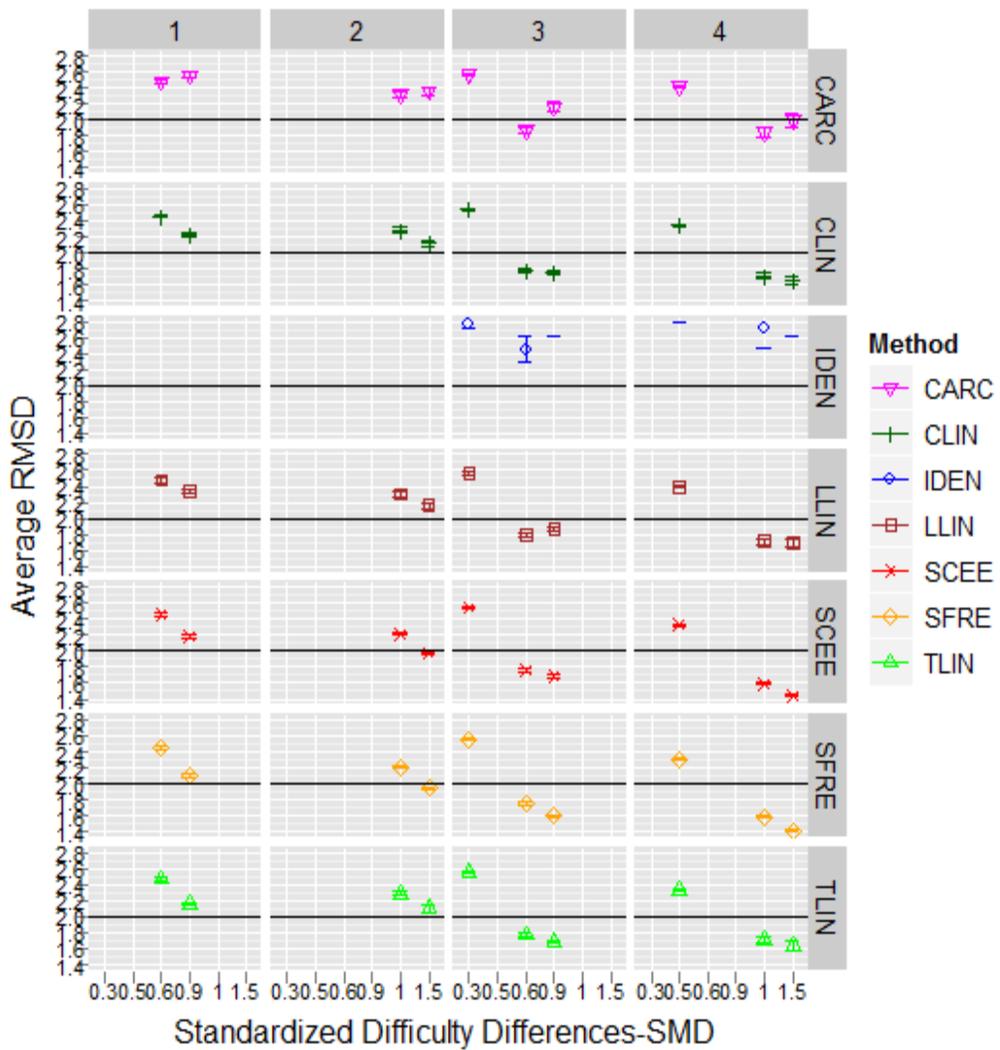


Figure 4.15: Distribution of the 95% CI of the Average RMSD as a Function of Discrimination (a-ratio) for the Various Methods by Studies Combinations

as a function of a-ratio, computed from Table 4.8, is illustrated in Figure 4.15. The panels with no plots in the figure do not form part of the design and can be ignored. The panels with only the horizontal solid line indicate that the 95% CI of the av.RMSD due to the equating methods for this set of conditions lies beyond the range of 1.2 to 3.0 used in this chart. This range excludes outlier conditions that would otherwise reduce the readability

of the plots in Figure 4.15. The horizontal solid line at the 2 point tick mark is a demarcation line between Studies 1 and 2 from Studies 3 and 4. Recall that the test forms in Studies 1 and 2 are half the length of those in Studies 3 and 4.

The following remarks are in order here. First, a higher a-ratio will improve the accuracy of all the equating methods with the exception of the CARC and the IDEN method. An a-ratio of 1.5 will guarantee that the equating results will be the most accurate. Second, doubling the test length has the merit of driving the total equating errors further down. In fact, the av.RMSD values go below the 2.0 mark, in Studies 3 and 4 where the tests are made of 60 items. However, this is true only if the a-ratio is 1.0 or higher. Third, an a-ratio of 0.5 will not be enough to bring the av.RMSD values below the 2.0 point mark even in the presence of a long test. Fourth, the larger base form average item discrimination in Study 2 compared to Study 1 (or Study 4 compared to Study 3) consistently drives down the overall equating error making the equating results for all the equating methods more accurate. In other words, increasing both the test length and the average item discrimination of the base form reduces the total equating error for all but the IDEN equating method.

The ranking of the equating methods in order of decreasing accuracy, ie., SFRE, SCEE, TLIN, CLIN , LLIN , CARC and IDEN, which is pretty typical based on the av.RMSD is no longer appropriate for all levels of the a-ratio. This ordering is true only when the a-ratio is 1.0 or 1.5 and most particularly when the base form discrimination is 1.0 as in Studies 2 and 4. If the base form discrimination is 0.6 as in Studies 1 and 3, (instead of 1.0 as in Studies 2 and 4), one might argue that the ordering is not as

applicable because all the NEAT methods are virtually equally accurate. If in addition, the test consists of only 30 items, as in Study 1, the CARC will join the list of NEAT methods.

An important revelation appeared when the base form discrimination a -ratio is 0.5 and the test is relatively short as in Studies 1 or 2. A reordering based on the ranks of the methods was obvious from the data in Table 4.8. In the case of Study 1 the SFRE and SCEE are very different from each other. They are no longer the most accurate among all methods and do not mimic each other as was the case in the analysis of the other factors. The same can be said about the CLIN and TLIN methods. These methods arranged themselves as they would typically when the measure of interest is the $av.BIAS$. In addition, the LLIN method became the most accurate of the NEAT methods with $av.RMSD$ of 3.25. The CLIN and SCEE methods were less accurate with $av.RMSD$ values of 3.34 and the SFRE and TLIN produced total equating errors of 3.43. The SE for the $av.RMSD$ is only .01 indicating that these differences between the methods are not trivial.

The same pattern is also present in the other studies where the a -ratio is .5 but the differences between the methods are not as drastically large. This is most likely a situation where the total equating error consists mainly of errors due to equating bias than from all other sources combined. For the same condition when the a -ratio is 0.5, in Study 3 there is virtually no difference among all the equating methods except the IDEN. In Study 4 for the same a -ratio of 0.5, the SFRE becomes the least biased and the LLIN is the most biased of the NEAT methods. The main conclusion from this is that the position

of the LLIN among the NEAT methods is a good indicator of the effect of the interaction of the a-ratio and the base form conditions contribute to the bias and the total equating error of the equating results.

It is important to note that the CARC outperforms all the other methods under the extreme conditions when the a-ratio is 0.5 as in Studies 1 and 2. In other words when the average item discrimination of the alternate forms is 0.3 or 0.5, the CARC is the most accurate. However, since in the previous analyses it tends to be very biased, this seems to confirm the hypothesis described above that the under these equating conditions (low a-ratio and poor base form conditions as in Study1) the total equating error consists mainly of errors due to equating bias than from all other sources combined.

Question 2: Interchangeability of Equating Results and av.RMSD

Figure 4.16 shows the effect of the discrimination ratio of the alternate form to the base form on the interchangeability of the equating results among the equating methods with respect to the av.RMSD. Scanning the plots column-wise indicate that there is no consistency in the relative separations among the equating across the three levels of the a-ratio. Furthermore, the relative separations among the equating methods and the base form characteristics interacts with the a-ratio to the extent that the clustering patterns are more inconsistent than observed in the previous cases.

First, when the discrimination ratio is 1.5, and the base form is adequately discriminating (a=1.0), the SFRE and SCEE will produce equating results which can be considered to be exchangeable with each other. Similarly the TLIN and CLIN methods will form another

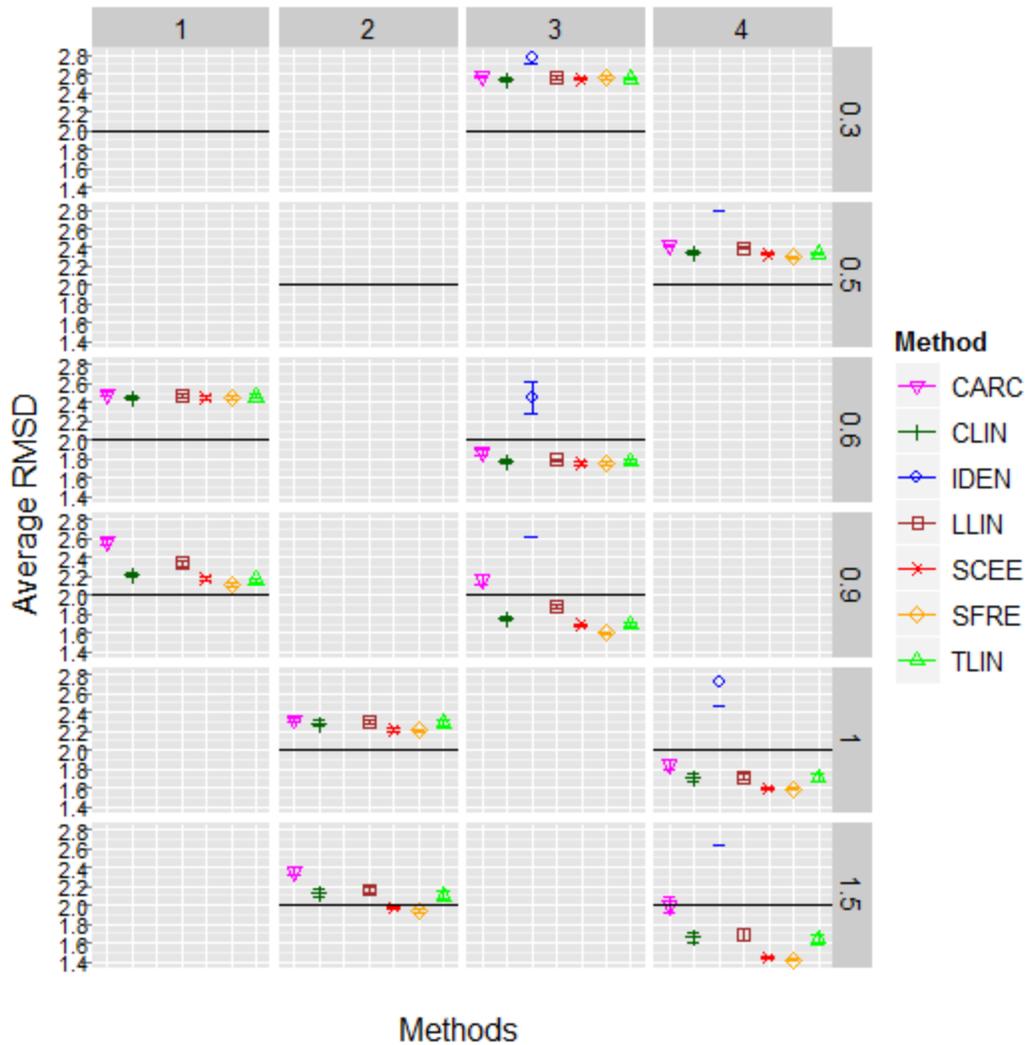


Figure 4.16: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various Form Discrimination (α -ratio) by Studies Combinations

cluster and the LLIN method might be considered part of that cluster relative the SFRE/SCEE cluster.

Second when the discrimination ratio is 1.5, and the base form is not adequately discriminating ($\alpha=0.6$) as in Studies 1 and 3, all the equating methods with the exception

of the SCEE and TLIN, might be considered practically non-exchangeable with one another (see the third row for Studies 1 and 3 in Table 4.8 for evidence of this finding). The SFRE and LLIN will distance themselves in opposite directions from the rest of the NEAT methods. The SFRE will be the most accurate and the LLIN will be the least accurate. The TLIN and SCEE are now more exchangeable with each other than the clusters they formed with the SFRE and CLIN in the previous situations.

Third, when the discrimination ratio is 1.0, and the base form average item discrimination is 0.6, all the equating methods with the exception of the IDEN and CARC are equivalent with each other regardless of test length. This might falsely signal that these conditions are optimal equating conditions. This notion may be reinforced by the fact that when the test is shorter the CARC joins the NEAT methods. With nearly 6 methods producing the same av.RMSD, it is hard to argue that these conditions (30 items, low base form discrimination, 0.6, alternate forms that are equally discriminating as the base forms) are not optimal and yet they are not. Some elaboration on this finding is provided further on.

This is a situation that might happen in small sample equating because it is quite possible that because of the lack of sufficient data (due to very low participation, fewer administrations or other practical considerations) the quality of the base form might be suspect. Future alternate forms built from such a base form may result in the kind of agreement between the methods when in fact, there is every indication that when the base forms are of adequate quality (1.0) and the alternate are at least equal in discrimination, the clustering into the “non-linear smoothing and the linear groups is to be expected.

Fourth when the discrimination ratio is 1.0, and the base form average item discrimination is 1.0 (as in studies 2 and 4), the NEAT methods are split into two clusters: the “non-linear smoothing” cluster and the “linear” cluster where the smoothing cluster is always substantially more accurate than the linear cluster (see the second row of Table 4.8, referring to the results for Studies 2 and 4, as an illustration of this finding). In the event that the test is the shorter version, the CARC method would be a potential addition to the “linear” cluster.

Fifth, when the discrimination ratio is 0.5 and the base form is short in length the formation of the clusters of the NEAT methods follow the trends that are more typical of av.BIAS where LLIN is more accurate than the CLIN/SCEE pair which in turn is more accurate than the SFRE/ TLIN pair. These clustering are particularly significant in the case when the alternate forms consists of the short version of the test and their average item discrimination are 0.3 or 0.5 which is very low by any standard. The typical SFRE/SCEE and TLIN/CLIN never occur under any circumstances when the a-ratio is .5. In other words the equating results are exceedingly inaccurate because of very large bias.

Effect of Test Length on Average BIAS

Recall that the test forms in Studies 1 and 2 are half the length of those in Studies 3 and 4. The 95% CI of the average BIAS for the various methods with respect to test length, presented in Table 4.9 and displayed in Figure 4.17, indicates that the shorter tests lead to slightly more biased equating results than when the tests are longer. The LLIN method is the least biased among all methods. The patterns that occurred in the

Table 4.9: Effect of Test Length on Average BIAS and SE of Equating Methods for Studies1-4

<i>Study</i>	<i>Length</i>	<i>av.BIAS</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
1	30	Mean	-0.11	-0.12	-0.18	-0.18	-0.03	0.05	0.90
		SE	0.01	0.01	0.01	0.01	0.01	0.02	0.10
2	30	Mean	-0.16	-0.18	-0.22	-0.21	-0.10	0.13	1.26
		SE	0.01	0.01	0.01	0.01	0.01	0.03	0.12
3	60	Mean	-0.09	-0.10	-0.14	-0.14	-0.05	0.58	0.93
		SE	0.01	0.01	0.01	0.01	0.01	0.02	0.10
4	60	Mean	-0.08	-0.10	-0.14	-0.12	-0.04	0.71	1.24
		SE	0.01	0.01	0.01	0.01	0.01	0.03	0.12

earlier analyses of the factors impacting the bias are repeated in this case as well. The LLIN remains the most accurate method followed by the CLIN, SCEE, TLIN, SFRE, CARC and least most biased is the IDEN. As in previous analyses the pairings between the SFRE and TLIN and the CLIN and SCEE methods are robust to variations in test length. But, a comparison of the av.BIAS of Studi 1 to Study 3 or Study 2 to Study indicates that the equating results within each cluster appear to be slightly more exchangeable when the base form average discrimination is small (0.6).

With regards to the CARC it is way much more biased on the 60 item test forms than the 30 item tests. In this study it is almost on par with the NEAT methods for the 30 item test except that it tends to be positively biased whereas the other NEAT methods are in the opposite direction. On the 30 item test, it actually is slightly less accurate and less stable than the LLIN method but it is more accurate than all of the other NEAT methods.

However, equating the 60 item tests causes the CARC to be much more biased as it is not even within the -0.5 to 0.5 range used to compare other NEAT methods.. As for the IDEN, the data in Table 4.9 shows that it is simply the most biased for both test

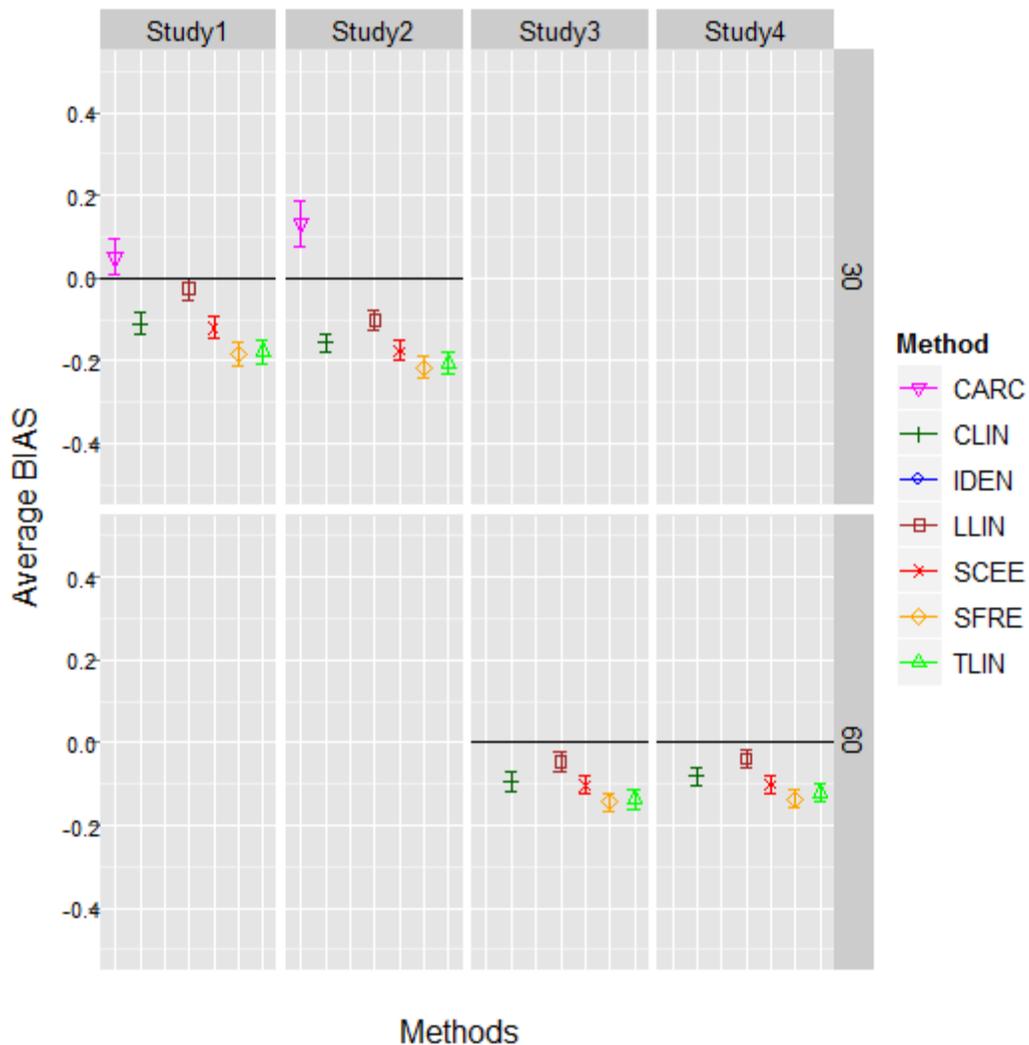


Figure 4.17: Distribution of the 95% CI of the Average BIAS as a Function of Methods for the Various Test Length by Studies Combinations

lengths. The relatively small bias of the CARC on the 30 item test is consistent with the fact that the aggregation of the equating bias across all the possible sampling and test characteristic conditions neutralizes whatever positive and negative biases of the equating transformations. The same comment applies to the LLIN method. The effect of the base form average item discrimination on the av.BIAS is not consistent across the two

different test lengths. The shift in the base form average item discrimination from 0.6 to 1.0 significantly drives up the av. BIAS of the NEAT equating methods on the 30 item test forms but has virtually no effect on the 60 item test forms. In brief equating short test forms will result in large amounts of bias. Longer test forms result in more accurate results and are less susceptible to the effect of average base form discrimination.

Effect of Test Length on Average RMSD

Table 4.10: Effect of Test Length on Average RMSD and SE of Equating Methods for Studies 1-4

<i>Study</i>	<i>Length</i>	<i>av.RMSD</i>	<i>CLIN</i>	<i>SCEE</i>	<i>SFRE</i>	<i>TLIN</i>	<i>LLIN</i>	<i>CARC</i>	<i>IDEN</i>
1	30	Mean	2.67	2.66	2.66	2.68	2.69	2.74	3.18
		SE	0.03	0.03	0.03	0.03	0.02	0.02	0.04
2	30	Mean	2.48	2.40	2.40	2.48	2.49	2.54	3.26
		SE	0.02	0.03	0.03	0.03	0.02	0.02	0.06
3	60	Mean	2.02	2.00	1.97	2.00	2.07	2.20	2.68
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.05
4	60	Mean	1.90	1.79	1.77	1.89	1.93	2.09	2.85
		SE	0.02	0.02	0.02	0.02	0.02	0.02	0.07

Table 4.10 shows the av. RMSD and the associated SE of the various equating methods for each of the four studies. It is clear that doubling the test length caused a drop of about half a point in the overall average RMSD. The av. RMSD of the CLIN method under Study 1 is 2.67 whereas the av. RMSD under Study 3 is 2.02. Similarly the av. RMSD of the CLIN under Study 2 is 2.48 whereas under Study 4 it is only 1.90. This result is consistent across all the equating methods. However, except with respect to Study 1, the SFRE seems to have the smallest av. RMSD among all the methods followed

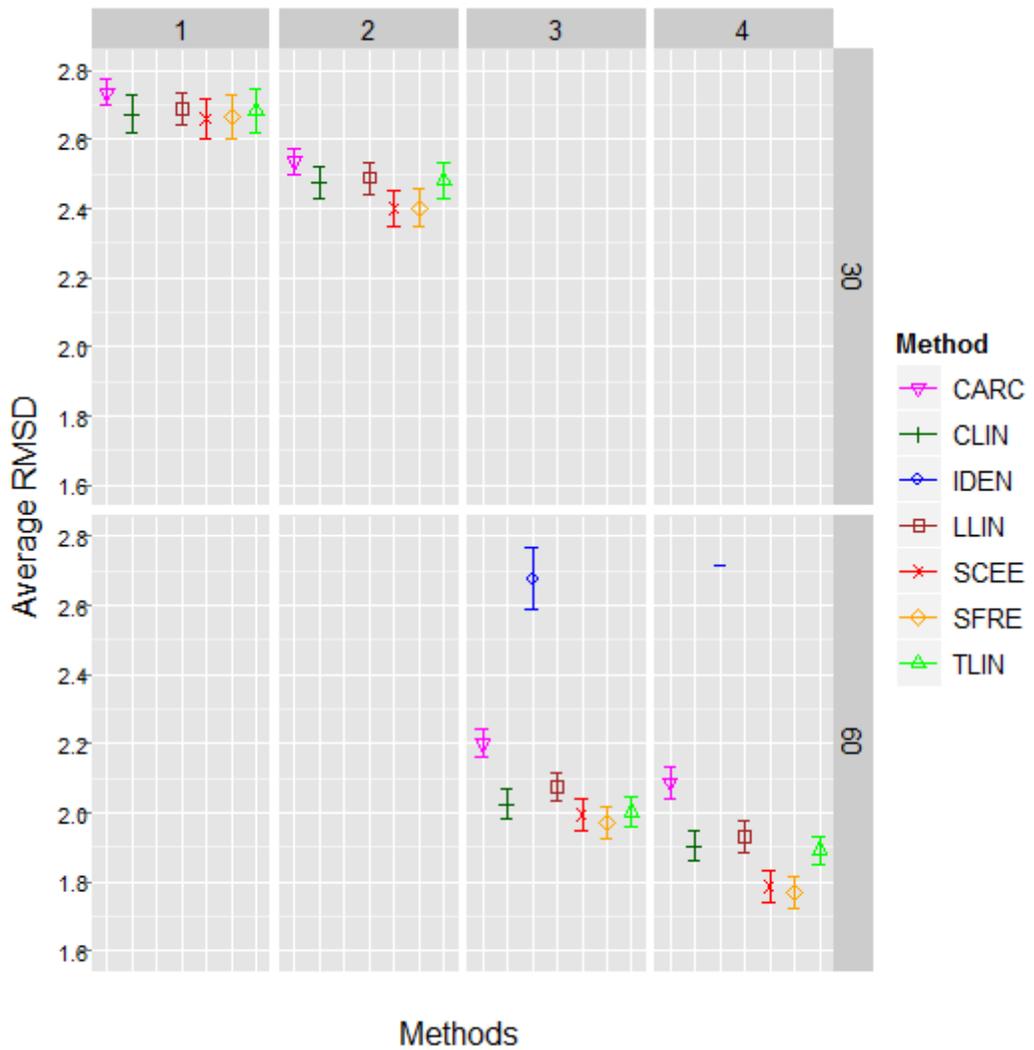


Figure 4.18: Distribution of the 95% CI of the Average RMSD as a Function of Methods for the Various Test Length by Studies Combinations

very closely by the SCEE. Among the linear methods, the CLIN and TLIN seem to parallel one another closely whereas the LLIN is generally the one which results in the largest av.RMSD among the NEAT design methods. The CARC for its part is consistently less accurate than the NEAT methods but is better than the IDEN method in general across both test lengths

In terms of the accuracy of the equating methods relative to one another, the SFRE is the most accurate followed by the SCEE, TLIN, CLIN, LLIN, CARC and IDEN. However, the overall differences among the NEAT methods are almost ignorable in Study1 where the forms consist of only 30 items and the base form average item discrimination of 0.6 is relatively low. In other words the methods are virtually exchangeable under these conditions. In Study 2, the SFRE/SCEE are practically identical and the same is true for the linear (CLIN, TLIN, LLIN) methods. The separation of these NEAT methods into these two clusters can be attributed to the base form average item discrimination which is now 1.0 instead of 0.6 as in study 1. The separation of the methods in Study 3 is less clear than in Studies 2 or 4, but in a general way, the LLIN method is no longer part of the clusters. The SFRE and SCEE and TLIN might be considered as one cluster or the CLIN, TLIN and SCEE might be considered as another cluster. But what is more certain is that within the limits of random error all the four methods could have been practically exchangeable with one another. The results for Study 4 confirms the trend noticed in Study2 where the SFRE and SCEE bond together and the TLIN and CLIN form the other pair that produce exchangeable equating results.

In short, the same patterns and trends observed of the impact of test length on the average RSMD of the various methods is similar to those observed with respect to other factors described in the earlier sections. However, this section clearly shows that the base form average item discrimination is a factor not to be ignored when it comes to the way it affects the (RMSD) overall equating error of the various methods and the extent to which their equating results are exchangeable. It is easy to see from Figure 4.18 that the longer

test has a very significant effect on the total equating error on all but the IDEN method. Together a longer and more discriminating base form will accentuate the differences between the equating methods..

Final Note

It is important that to underscore that the average bias, $\frac{1}{N} \sum_{j=1}^N d_j$, computed was not centered at zero after the results were collapsed over the various combinations of conditions. The simple and straight explanation for these observations is related to the STD. Under ordinary circumstances, bias should be zero in expectation. However, in this study, the expected bias is zero for all equating methods only when the groups are equally able (STD =0). Under all other circumstances the av.BIAS is not zero because the alternate form groups are always more able (STD=0, 0.05, 0.10, 0.25) than the base form group. Actually the statistic was always negative, in particular with respect to the NEAT methods. This is a clear indication that the equated scores were almost always underestimated relative to the “true” scores. In addition, the fact that the alternate forms that they were administered were on average easier (SMD= - 0.75, -0.50, -0.10, 0, 0.25) than the base forms contributed to making the av.BIAS even more negative.

CHAPTER V

CONCLUSIONS AND DISCUSSION

Accuracy Summary

The first research question seeks to determine the effects that variations in sampling and test characteristics have on the accuracy of the various equating methods whereas the second research question addresses the issue of how these variations affect the interchangeability of the equated scores among the equating methods. This chapter provides a summary and a synthesis of the detailed descriptions of the findings exposed in chapter IV related to each question. These are accompanied by some of the implications related to the findings.

Some of the findings conform to already well established facts. Others may be considered as unexpected results and still some have raised more questions than answers. There is strong evidence of some recurring themes that occurred for some conditions that are in agreement with previous studies and others which are less clear. At the same time as this study attempted to answer the research questions it also provided some insight into how the findings could be used for diagnostic purposes on the basis of the accuracy measures.

Accuracy and Sample Size

First, the most important result about the impact of sample size is that it has no effect on the av.BIAS of the equating methods. This is not to say that any sample size

will do. In fact, the stability of the equating results are consistently smaller when the sample size is 50 or less compared to the sample size of 200 or more for all except the CARC and IDEN methods. This study has shown that in the universe of all possible combinations when the sample size is 50 the risk that the estimated equating results deviates from the “truth” is likely to be nearly 50% (based on the standard error) greater than if the sample size is 200 or more. Actually this is to be expected given that sample size 25 or even 50 are far too small (or extreme) to ensure that the equating results do not vary substantially along the spectrum of all sampling and test characteristic conditions (which also include other extreme levels) with which they are associated.

On the other hand, the effect of sample size on the av.RMSD is different. The accuracy of the equating methods continuously drops almost imperceptibly as the sample size increases from one level to the next. Nonetheless the drop is steady even if it is slow unlike the bouncing around with sample size that occurred for the av.BIAS. The result may seem surprising at first because the expectation is that the drop should have been much more drastic from one level to the next and this did not happen in this study. One possible explanation for this much smaller than expected drop in the av.RMSD is that much of the variability that is associated with sampling is minimized. The design of the study is such that sampling error is minimized right from the very first steps of data generation. The subsequent steps of equating to the truth, computing the RMSD over examinees followed by averaging these RMSD’s over 10 replications and over each level of sample size minimized the effect of sampling error between extremely and not so small samples.

Overall the most important point about sample size is that it affects the total equating error (RMSD) as much as it has no effect on the bias of the equating methods and this result is not a surprise. Equating that involves samples of size 50 or less are very susceptible to relatively larger inaccuracies both in terms of the total error as much as in the imprecision of the estimates of the bias. When in doubt or very little is known about the sampling or test characteristics a minimum sample size of 200 may be the answer for fairly accurate and reliable. It is important that great caution is used when equating samples of size 100 or smaller because of the greater risks of inconsistent results if very little information is known about the test characteristics or group differences. Equally important is that these recommended sample sizes are only approximate and may vary with equating conditions and equating methods. These findings are not unlike those reported in previous studies by Skaggs (2005) and Kim and Livingston (2010).

Accuracy and Magnitude of Group Separation (STD)

Similar to sample size, the magnitude of group separation (STD) had different effects on the av.BIAS and the av.RMSD of the equating methods. As the groups become more dissimilar in ability the equating results of all the equating methods became increasingly biased and at the same time they became less reliable. However, if the overall equating error (RMSD) is the measure of accuracy of interest and not bias, then the effect of magnitude separation based on the av.BIAS may go unnoticed. In fact, the av.RMSD of the equating methods do not show any sign of varying with the STD. The accuracy and stability of the estimated equating results remained constant with changes in the magnitude of group separation. These two contrasting results are major highlights of

the effect of the STD on the accuracy of the equating methods. The immediate consequence of these two results is that the overall equating error (av.RMSD) will not reflect the magnitude of the group separation and yet the equating results are likely to be exceedingly biased if the STD is large. These results are remarkable to the extent that they were not anticipated and to the best of my knowledge there is no reference in the literature to that effect. This study explicitly showed the differential effect of group differences on the two measures of accuracy.

Overall, based on the results of this study the bias is a far more important criterion than the total equating error when it comes to deciding about the level of group separation that would not overly distort the equating results. Among the four levels of STD examined in this study group separation as large as 0.25 STD is definitely unacceptable. This result is consistent with the results reported by Petersen (2007), Wang et al. (2008). The recommendation based on this study is that group separation should not exceed 0.1STD and should be as close to zero as possible because even group separation of 0.05 may lead to biased results if other conditions are not optimal. The quote from Wang et al. (2008, pp.635) referring to group differences “ In test equating, mean differences between .05 and .1 are generally considered relatively large, whereas a mean difference of .25 is usually considered a very large difference” is more than appropriate in this case. One might argue that these remarks by Wang et al. are based on the total equating error and not bias whereas in this study the same conclusions are reached from the measure of bias. The reader is reminded that the conventional approach to computing total equating error reported in the literature includes two components, bias

and sampling error and that it is generally derived from large samples. As such when the random error is minimized through the use of large samples and reaches some asymptotic value any difference between equating methods in the total equating error can be attributed to the bias of the methods than to sampling error. Since group differences have the most impact on the bias, this explains why the conclusions about the STD based on the av.BIAS in this study are no different than the findings of Wang et al. This study has shown that differences in group ability is probably the most potent factor that affects the bias of the equating results and this is exactly what this study has shown.

Accuracy and Test Difficulty (SMD)

In contrast to the effects of the magnitude of group separation (STD), differences in test difficulty (SMD) has the opposite effects on the av.BIAS and the av.RMSD. The av.BIAS of the various methods (except the CARC and the IDEN) barely changed with varying SMD between the equated forms. Similarly the precision of the av.BIAS remained virtually constant across the various levels of SMD. The findings of this study showed that it would not be unreasonable to conclude that there is no association between the bias of the NEAT methods and differences in test difficulty.

However, the effect of the SMD on the total equating error (av.RMSD) is virtually constant for all the equating methods until the equated forms became exceedingly different in difficulty (i.e., -.50 or -.75). At these levels of SMD, there is a break down in the accuracy of the equating results that is related to the equating methods. The “non-linear smoothing” methods are systematically more accurate than the “linear” methods. This result is not unexpected because linear equating functions are not designed

for handling situations where there are large differences in test difficulty between the base form and alternate forms. However, more remarkable is the extent to which the equating results were nearly almost as accurate when the SMD was 0.25 as when there was no difference in test difficulty.

Overall, in stark contrast with the conclusions about the impact of the group differences, the results about the impact of test difficulty indicate that total equating error is a far more important criterion than bias in deciding about the level of test difficulty that would be reasonable for test equating. Of the five levels of SMD examined in this study any SMD level greater than 0.25 is definitely unacceptable. The equating methods in general are capable of adjusting differences in test difficulty that range between 0 and 0.25 SMD in absolute value.

Accuracy, a-ratio and Base Form Discriminations

The effect of the discrimination ratio of the equated forms to the base form is more complex but unlike the other factors (sample size, STD, SMD) there are no large contrasts in the equating results with respect to the av.BIAS or av.RMSD. In fact, the conclusions about the effect of the discrimination ratio based on the av.BIAS support those based on the av.RMSD.

Increasing the a-ratio has the direct effect of reducing the av.BIAS of the equating methods. The reduction is more drastic when the a-ratio changes from 0.5 to 1.0 than from 1.0 to 1.5. There is therefore, a higher risk of larger biased equating results when the ratio of average item discrimination of the alternate forms to the base forms is low. In fact, this study has shown that equating alternate forms that are substantially less

discriminating than the base forms can lead to relatively larger underestimation of the equating results. As much as possible the alternate form average item discrimination should be as close to if not higher than that of the base form.

Furthermore as stated in the previous paragraph the conclusions about the effect of the a-ratio on the av.RMSD support the conclusions about the effect on the av.BIAS. Low a-ratio (0.5) leads to relatively much larger total equating error than high a-ratio. There is a much larger improvement in the equating accuracy when the a-ratio changes from 0.5 to 1.0 than from 1.0 to 1.5. These results are not unexpected because the test information is literally increased by a factor of about 400% when the a-ratio changes from 0.5 to 1.0 whereas it increases by only 225% when the a-ratio changes from 1.0 to 1.5.

Equally important is the average item discrimination of the base form itself. This applies to the impact of the a-ratio on both the total equating error and bias. A base form which is not very discriminating will result in larger total equating error and more biased equating results than a more discriminating one. Indeed a base form with an average discrimination of 0.6 can be thought of as possessing only 36% the test information of a base form with average item discrimination of one. So to minimize the risk for biased equating results, two conditions need to be met: 1) avoid equating when the a-ratio is low and 2) avoid equating to poorly discriminating base form. The recommended a-ratio based on this study should be close to 1.0 or higher and as far as possible low discriminating base forms should be avoided.

These conclusions imply that differences between the equated forms will be reflected consistently in the bias and the total equating error of the various methods. They also imply that the test information of the alternate forms relative to the base forms is a very important factor. It appears that test information is a very sensitive indicator of differences between the base forms and the equated forms. In fact, because it capitalizes on both the av.BIAS and the av.RMSD it may be used as a complement to the SMD to assess the quality of a test.

Accuracy and Test Length

With reference to the impact of test length, this study has shown that it has the same consistent effect on the av.BIAS and the av.RMSD of the equating methods. Both measures indicate that the equating results become less biased and the overall equating error is substantially reduced when the sample size is doubled. Equally notable is the fact that in the universe of all possible conditions, the equating results are more stable in the case when the test is doubled in length. This conforms to the already well established notion that, everything else being equal, a longer test in general is likely to be more accurate and reliable than a short one both in terms of the total equating error or the bias of the equating results. Since this is a simulation study these results about the impact of test length are not unexpected.

Accuracy and Equating Methods

It also worth noting that there are marked differences between the equating methods when they are arranged in order of decreasing accuracy. The order associated with the av.BIAS is different than that based on the av.RMSD. In general, based on the

av.BIAS, the methods follow the sequence: LLIN, CLIN, SCEE, TLIN, SFRE, CARC and IDEN. Based on the av.RMSD the sequence followed is: SFRE, SCEE, TLIN, CLIN, LLIN, CARC and IDEN. Both sequences are practically unchanged across most sampling and test characteristic combinations included in this study. More on this subject follows in the next section.

Interchangeability Summary

This section provides a general summary of the effect of the various sampling and test characteristics on the interchangeability of the equating results among the equating methods. As in the case of the findings related to the first research question there are some recurring themes that were revealed from the analyses related to research question 2. The summaries refer to essentially to the NEAT methods and the CARC and IDEN methods are included only where appropriate as these two methods were exceptions in most cases. Similar to the inconsistency of the impact of the various sampling and test characteristics on the accuracy of the equating methods with respect to the av.BIAS and the av.RMSD, the impact on these measures by these same factors were equally inconsistent when the exchangeability of the equating results was compared across equating methods.

Interchangeability and BIAS

A key finding of this study is that the pattern of exchangeability among the methods with respect to the av.BIAS was virtually invariant across most of the combinations of sampling and test characteristics, except when there was no difference in

the ability ($STD = 0$) of the base form and the alternate form samples. The equating methods can be grouped into clusters that remained unaltered regardless of variations in sample size, the magnitude of group separation (other than $STD = 0$), variations in differences in test difficulty, or test length. The CLIN and the SCEE methods always paired together and the same applies to the SFRE/TLIN methods. In other words, the CE methods were always less biased than the PSE pair. In very rare instances these pairings did not occur. The most important and of greatest significance are those clusters that formed when the base form and alternate form samples are both equally able. Under these circumstances all the NEAT equating methods could be considered to be practically equally biased and equally stable for most intent and purposes.

One plausible explanation for the fact that the same clustering pattern occurred for most sampling and test characteristic conditions is that the bias due to group differences is a dominant factor even in the presence of other sampling or test characteristics. When the STD is equal to zero the equating methods did not have to do any adjustment between the base form samples and the alternate samples. Overall the ultimate effect is that group separation drives the bias of the equating methods. More than any other factor group separation may be the most important factor that affects the bias of the equating methods even if it is not the most important factor that drives the total equating error.

Conclusion. Overall the pattern of exchangeability among the methods with respect to the $av.BIAS$ was virtually invariant across most of the combinations of sampling and test characteristics, except when the equated samples were equally able.

Interchangeability and Total Equating Error (RMSD)

What can the patterns of interchangeability of the equating results among the various equating methods based on the total equating error tell us about the impact of the various sampling and test characteristics? This study has revealed that the methods that are likely to produce consistently exchangeable results can be categorized into different clusters that depend not just on sampling or test characteristics but also on their interaction with the base form. In general, except for Study 1, the NEAT methods can be categorized into three clusters: LLIN, SFRE/SCEE and TLIN/CLIN. These clusters join together or separate from one another to form new clusters depending on the sampling/test characteristics and or the base form conditions.

Overall the exchangeability of the equating results based on the av.RMSD is more similar between Studies 1 and 3 or between Studies 2 and 4 than between Studies 1 and 2 or Studies 3 and 4. This is so because Studies 1 and 3 or Studies 2 and 4 only differ in test length whereas Studies 1 and 2 or Studies 3 and 4 differ in the base form average item discrimination. It is to be noted that a longer and more discriminating base form accentuate the separation of the methods thereby making them less interchangeable.

Study 1: A Special Case. It is important to point out that the findings from Study 1 are different than all the others to the extent that all the three clusters- LLIN, SFRE/SCEE and TLIN/CLIN- stick together except under extreme sampling or test characteristic condition. All the NEAT methods are virtually exchangeable with one another under the worst base form condition (30 items and average item discrimination of 0.6) unless the equating condition is extreme, such as sample size of 25, a-ratio of 0.5 or

SMD of $-.75$. In addition, the CARC method is not much more or less accurate than the NEAT methods, though it tends to be more stable than them. These results were not anticipated because the base form is the least reliable ($\alpha 0.78$) and has the lowest anchor to total test correlation ($r = 0.81$) among all the four base forms used in this study and yet the NEAT methods are virtually exchangeable with one another. The formation of the clusters under Study 3 is not very different from Study 1 except that the LLIN method tends to distance itself from the other NEAT methods. The formation of the “linear” and “non-linear smoothing” clusters under Studies 2 and 4 are virtually identical and more in line with what would be expected.

One plausible reason for the formation of one big cluster in Study 1 may be due to the very poor base form conditions (short test length and low discrimination) and the associated alternate forms which are relatively poorer than in the other studies. It is very likely that because of the low base form average item discrimination, the base form sample does not discriminate very well between the examinees. This would make the alternate groups which take forms which are lower in discrimination (a-ratio 0.5) even more homogeneous. The alternate form groups which take forms which are as poorly discriminating as the base form (a-ratio of 1.0) will be as equally homogeneous. Only those groups taking the more discriminating alternate forms (a-ratio of 1.5) will maintain some degree of discrimination between the examinees that make up a group. However, the differences among the examinees in the groups taking the more discriminating alternate forms may be washed away when the groups are equated to the poor base form. As a result, the actual differences between the groups in Study 1 are not reflected in the

equating results. In other words there is apparently far more homogeneity among the groups in study 1 and the base form sample than in the other studies. This homogeneity overrides the actual ‘true’ ability differences between the alternate and base form samples making the difference between their anchor on both tests minimally small. As the difference on the anchor between an alternate from group and the base form becomes very small all the equating methods tend to give the same results (Kolen, 1990).

Sampling Characteristics. With respect to sampling characteristics, the exchangeability among the equating methods based on the av.RMSD may be reduced to essentially two main clusters: the “linear” cluster and “non-linear smoothing” cluster for the longer test (Studies 2 and 4). The “linear” cluster consists of all the linear methods and the SFRE/SCEE forms the “non-linear smoothing” cluster. In Study 3 the clustering is a hybrid of the clusters of Study 1 and Study 2. Both the LLIN method and the “non-linear smoothing” pair are not fully exchangeable with the TLIN/CLIN pair. The LLIN method is relatively less accurate and slightly more stable than the TLIN/CLIN pair whereas the SFRE/SCEE pair is slightly more accurate than both. It would not be unreasonable to argue that the clustering in Study 3 is nearly the same as in Study 1 except that the LLIN method is less exchangeable with the other methods, because in many instances the “non-linear smoothing” cluster and the TLIN/CLIN pair are very close to each other. For the sake of simplicity the NEAT methods in Study 3 can be considered to be split in two clusters: LLIN and the TLIN/CLIN/SFRE/SCEE.

The exchangeability of the equating results among the equating methods in study 4 relative to the sampling characteristics are virtually similar to the formation of the

same two clusters that formed in Study 2. The only major difference is that this time the “linear” clusters and the “smoothing” clusters are now fully distinct from another. This difference may be attributed mainly to the effect of doubling the test length. The equating results among the equating methods within these two clusters are virtually exchangeable with one another but nearly almost completely not exchangeable across clusters.

The summary provided in the previous paragraph refers only to the typical clustering patterns associated with sampling characteristics and does not make any mention of the exceptions or atypical situations. The atypical situations arise mainly when the level of the factor of interest is unrealistic or extreme. For example, for sample size 25, the equating results of the LLIN method was substantially different for the other NEAT methods in study1. Nothing in particular occurred with respect to STD that did not reproduce the clustering patterns of described above.

Test Characteristics. The interchangeability of the equating results based on the av.RMSD among the equating methods with respect to the test characteristics is different than based on sampling characteristics. The cluster formations within a study are not always independent of the variations in the level of the test characteristic of interest and/or the base form conditions.

Test Difficulty Differences (SMD). In the case of the SMD, a major split in the NEAT methods occurred when the difference in test difficulty exceeded 0.25. The clusters then consisted mainly of the “linear” and the “non-linear smoothing” methods. These two clusters were essentially exchangeable with one another when the difference in

form difficulty did not exceed 0.25. In fact, when there is no difference in form difficulty between the base form and the alternate forms the IDEN and CARC methods would join the NEAT methods and all the equating methods are virtually exchangeable with one another. However, the CARC tends to be more exchangeable with the other methods when the base form is not very discriminating and/or the test is short in length. The IDEN for its part is mostly exchangeable with the other methods when there is no difference in difficulty between the forms. Under most other conditions these two methods are not exchangeable with the other methods.

The split of the NEAT methods into the linear and “non-linear smoothing” clusters when the SMD exceeded .25 is not unexpected because “linear” methods are more appropriate for equating forms that are not very different in difficulty from one another and the nonlinear methods like the SCEE and SFRE are more appropriate for equating forms that largely differ in difficulty. The most unanticipated result was that at 0.25 SMD the exchangeability had not changed much compared to when there is no difference in difficulty between the equated forms. But the split in the clustering when the SMD changed between .25 to -.75, suggests that somewhere in between these two levels there might exist a threshold SMD value below which the NEAT methods remain exchangeable with one another and beyond which the NEAT methods split into the two clusters. However, the base form condition must also be factored into the equation because in study1, except when the STD was -.75, all the equating methods (excluding the IDEN when the SMD was 0 or -.1) remained solidly exchangeable with one another.

a-ratio and Base From Discrimination. The situation based on the av.RMSD with respect to the a-ratio is somewhat similar to that of the SMD to the extent that the level of the a-ratio affects the exchangeability of the equating methods within a study but there is also a much stronger interaction with the base form conditions that affect the exchangeability across the studies. None of the “smoothing” and or “linear” clusters occurred when the a-ratio was 0.5. Instead when the a-ratio was 0.5 and the test is short in length, as in studies 1 and 2, the SFRE and TLIN methods were more exchangeable with one another like they would, based on av.BIAS. Similarly the CLIN and SCEE methods were more exchangeable instead of CLIN and TLIN. The LLIN method was more accurate than either of these two pairs and the ranking based on accuracy was no different than would be expected when the measure of accuracy is the av.BAIS than av.RMSD. This is a very special situation and can be considered an anomaly. One plausible explanation for this occurrence is that when the discrimination ratio of the alternate form is half that of the base form, and the test is short the risk of obtaining equating results that are very strongly biased is very high. Doubling the test length as in Studies 3 and 4 corrects the problem to some extent and makes the equating methods more exchangeable.

The situation is entirely different when both the base form and the alternate forms are equally discriminating. If the equated forms are equally low in discrimination (i.e., a-ratio is 1.0 and the base form discrimination is 0.6), all the NEAT methods are virtually exchangeable with one another. But if the base form discrimination is more discriminating (1.0 instead of 0.6) the NEAT method splits into the “non-linear smoothing” and “linear” clusters.

Furthermore the clustering patterns observed when the a-ratio is 1.5 and the base discrimination is 1.0 is similar to that of the case when the a-ratio is 1.0 and the base form discrimination is 1.0, except that the LLIN method is no longer exchangeable with the other two linear methods. In the case when the a-ratio is 1.5 and the base form discrimination is 0.6, a peculiar arrangement occurred. Both the “non-linear smoothing” and TLIN/CLIN clusters split up. The SFRE was much more accurate than SCEE method, which in turn became more exchangeable with the TLIN which in turn dissociated itself with the CLIN method. In other words the SFRE/SCEE and TLIN/CLIN clusters which is prevalent across most of the combinations examined in this study, changed to SFRE, SCEE/TLIN and CLIN where they are arranged in order of decreasing accuracy. To summarize, both the base form discrimination and the a-ratio are important factors that affect the interchangeability of the various methods.

Test Length. There is clear evidence that based on the av.RMSD, test length has a direct effect on the degree of exchangeability of the NEAT methods relative to one another. In addition poor base form conditions tends to make all the NEAT methods virtually exchangeable with one another. As the base forms become more discriminating, the NEAT methods split into the “linear” and “non-linear smoothing” clusters. Increasing both test length and the base form discrimination accentuates the separation between the “linear” and “non-linear smoothing clusters, but within a cluster, the methods are practically completely exchangeable whereas the two clusters are stretched further apart.

Conclusion. Overall based on the av.RMSD, sampling characteristics do not affect the interchangeability of the equating results among the NEAT equating methods but test characteristics are more important factors in that respect. In particular, large differences in difficulty or extreme a-ratio of the alternate forms to the base forms have the largest effect on the interchangeability of the equating methods.

Recapitulation and Recommendations

Equating Conditions

The recommended sample size for successful equating should be about 200 or more if limited information is available in particular about group differences. A sample size of 50 or less might be problematic because of the high risk of unstable equating results. It would be best to avoid any equating of forms if the sample size less than 50. For the magnitude of group separation (STD) the recommendations are as follows: 0 is best, 0.05 may be acceptable, 0.1 acceptable under some conditions, 0.25 to be avoided at all cost. For test difficulty differences, SMD value of zero is best, -0.1 or -.25 are acceptable but -.5 or -.75 are too extreme and should be avoided at all cost. The smaller the SMD values the better. In terms of the a-ratio, 0.5 is unacceptable and should be avoided at all cost because of the high risk of both extremely biased results and relatively very high total equating error, 1.0 is acceptable, 1.5 is best. Needless to say, longer test forms are highly recommended. Put together it would be wise not equate marginal-to-poor quality tests with small samples. In fact, the same is true for large-sample equating.

Equating Methods

Arranged in order of decreasing accuracy based on the av.BIAS the equating methods follow the sequence: LLIN, CLIN, SCEE, TLIN, SFRE, CARC and IDEN. Two points are worth noting here. First all the NEAT methods are virtually equivalent when the equating samples are equally able. Second, the LLIN method stands out as the most consistently accurate method under most equating conditions. Contrary to all the other methods, it remained consistently centered on the zero bias line and is virtually robust to variations in different equating conditions (SMD, test length, sample size, or STD) except when the a-ratio of the alternate form to the base form is too low (0.5). This is probably among one of the most important findings of this study, given that the other NEAT methods are generally considered more popular since they are based entirely on observed scores. The LLIN method unlike the other methods makes use of “true” scores in estimating the equating function. As such it should be less biased than the other methods and this study showed just that.

In order of decreasing accuracy based on total equating error (RMSD) the equating methods follow the sequence: SFRE, SCEE, CLIN, TLIN, LLIN, CARC and IDEN. Exceptionally the IDEN and CARC performed (more accurate and more stable) consistently better than the NEAT methods when the equated forms are equal in difficulty. However, these two methods tend to be very biased with longer tests. It is to be noted though that the lower av.RMSD for the smoothed methods may be due to a reduction in the residuals but in the process the smoothing error may increase which in other words causes the bias to increase. A lower av.RMSD by no means guarantees that

the equating results are necessarily unbiased. A lower av.RMSD may be only a necessary condition and is only better if the av.BIAS is zero. Otherwise it indicates consistent bias. As such the LLIN method is undisputedly the best method as it is the least biased of all the methods.

To conclude the remarks reproduced from Chapter II (p. 35) of this study of the following quote from van der Linden (2010) could not be more appropriate in this case:

...bias is the primary criterion for evaluating the success of an equating. After all, equating is an attempt to remove the bias in the score on the new test form as an estimate of the score on the old form due to scale differences. A focus on the standard error of equating prevents one from noticing any remaining bias in the equated scores, or even possible new bias added to them in the equating process. (p. 21).

Base Form Conditions

Increasing test length reduces the bias and total equating error and improves the stability of the equating results. On one hand, the effect of increasing the average item discrimination of the base form on the bias of the equating methods is confounded with test length. If test is short is length, the increase in discrimination produces more biased (av.BIAS) results, whereas for the same increase in discrimination doubling the test length reduced the bias. On the other hand, the effect of increasing the average item discrimination of the base form on the total equating error (av.RMSD) of the equating methods is not affected by test length. The total equating error is reduced significantly as the base form average item discrimination is increased.

In terms of exchangeability based on the av.RMSD, increasing both test length and the base form average item discrimination forces the equating methods into clusters

of interchangeable methods. The two most common sets are the “non-linear smoothing” and “linear” clusters or the “non-linear smoothing”, TLIN/CLIN and LLIN clusters. Such clustering is true for the SMD only when the test difficulty differ by more than .25 SMD. Under smaller SMD conditions all NEAT equating methods are virtually exchangeable with one another.

The impact of the base form conditions on the interchangeability of the equating methods is more tractable. The formations of cluster and the relative separation from one another are nearly constant and virtually do not vary with the base form conditions. The NEAT methods are split into three clusters: LLIN, CLIN/SCEE and SFRE/TLIN. The only exception occurs when the base form group and the alternate form group are nearly or equally able. Under such conditions all the equating methods are nearly equally biased and the clusters do not always form as described. Overall the base form conditions impact the total equating error and the bias of the equating methods differently. This leads us to the next section where I try to provide some explanations to the results and conclusions reached.

Final Comments and Partial Explanation of Some of the Results

Other than a few studies in the literature on large samples where a “true” equating function is used as the criterion function to evaluate classical test theory equating methods, this study may be the only to have used such a criterion with small samples. In addition, in this study every attempt has been made to minimize as much sampling error as possible and to control the measurement precision. It certainly presented a novel

approach for small-sample equating and expanded on the test characteristic conditions investigated by Wang et al., 2008. In fact, the methodology, equating accuracy, and equating stability criteria used in this study were quite different than the conventional approaches—based on resampling methods—that seem rather prevalent in the literature. Specifically, this study defines equating error as the difference between the equated examinee scores and the known “true” scores based on an IRT model (assuming that the data fit the IRT model). The major weaknesses of this study one can argue is that differences between examinees along different intervals of the raw score scale are ignored in the computation of the accuracy statistics. But is it not that population invariance is one of the requirements for equating? In other words, the equating function used to link the scores of the alternate form to the base form should be the same regardless of the choice of (sub) population from which it is derived (Petersen, 2007).

On one hand, most of the findings of this study conform to or confirm some well known and established results. This study does provide some insight into the behavior of the equating methods. Some has been established over time or others from a theoretical stance but still others remain to be demonstrated empirically to be true.

First, the long standing reference to the equipercentile equating method as the “gold” standard is an example in case. This study did not specifically examine the equipercentile method because it is not conducive for small sample size equating. Instead, assuming the SCEE is a proxy for the equipercentile method, the study confirmed its superiority over the PSE methods. For the most part it is virtually nearly equivalent to the CLIN method and certainly consistently more biased than the LLIN method regardless of

equating conditions. The CARC and the IDEN methods are generally the most biased and also result in the largest overall equating error.

Second, the total equating error was the decisive factor in determining the level of test difficulty beyond which equating would not be appropriate whereas the measure of bias was the decisive factor in determining the threshold for group differences. To my knowledge I am not aware of any studies that have shown these results in such unequivocal way. In retrospect, these findings are reasonable to the extent that variations in group differences are expected to have more impact on the bias than on the total equating error to the degree that sampling error was minimized as in this study.

Third, the bias of the NEAT methods relative to each other did not come as a surprise. Based on the measure of bias the equating methods did perform as expected or as reported in the literature (see Kane, Suh, Mroch, & Ripkey, 2009a; Kane, Suh, Mroch, & Ripkey, 2009b; Suh, Mroch, Kane, & Ripkey, 2009; Mroch, Suh, Kane, & Ripkey, 2009). The LLIN method is less biased than the PSE methods and the CE class of methods are expected to fit somewhere in between them. The performance of these methods based on the total equating error can be considered to be less informative because it includes the variability of the equating error which to large degree is influenced by such factors as sample size or test length or test reliability. More significant are the actual differences between the equating methods themselves which are more ‘truly’ represented by the av.BIAS than the av.RMSD. As such even if the trend among the methods based on the total equating error did not follow the same pattern as on the measure of bias it should be too much of a concern.

Fourth a very interesting phenomenon was revealed. Under some conditions the NEAT methods are all virtually identical (such as in Study 1), or the linear methods of the NEAT group would be nearly interchangeable with one another and the two “non-linear smoothing” methods would cluster together. In other words, the NEAT methods would split along different lines based on the av.RMSD or av.BIAS. The clustering of the methods based on av.BIAS fell into classes organized by equating type (Levine, CE and PSE), whereas the clusters that formed based on the total equating error could be categorized into linear or non-linear types.

Fifth, von Davier, Holland, and Thayer, D. T. (2004b) actually showed that the CLIN method is a special case of the CE class of methods and that TLIN is the linear version of the frequency estimation method under the PSE class of methods for the NEAT design. The formation of the CLIN/SCEE and TLIN/SFRE clusters based on the av.BIAS and the consistently less biased estimates of the CE pair provides very strong support to the notion that the CE type is less biased than the PSE type of equating methods. For the most part, these results conformed with similar results from past studies by Livingston, Dorans, and Wright (1990), Sinharay and Holland (2010a), Puhan (2010) even if they were derived by a completely different approach in this study and involved the use of small samples, multiple conditions covering a several levels from the least to the most extreme.

Sixth, von Davier et.al (2004b) also provided two theoretical conditions under which the CE and PSE methods must produce interchangeable equating results. One condition requires that there are no differences in the base form and alternate form groups

(similar ability, $STD=0$) as measured by the anchor test and the other condition requires that the anchor be perfectly correlated with the base form and the alternate form. The findings of this study based on the *av.BIAS* suggest that there may be a close match to the first condition that the equated groups are of the same ability. The surprising result is that this is predicted on the basis of the total equating error rather than bias as determined in this study. In fact, based on the *av.RMSD* the CE and PSE methods never matched each other except in Study 1 where the base form was the worst among all four studies. In other words, the results based on the *av.RMSD* in study 1 matched the condition reported in von Davier (2008) that all the linear methods become exchangeable with one another as if there are no group differences ($STD=0$), regardless of whether they are from the CE, PSE or Levine class of methods. The cause for this “contradiction” where the results of this study is derived from the *av.BIAS* to similar results based on the *RMSD* as determined by von Davier is unclear and needs to be examined further.

In closing, a final comment is in order with respect to the definition of “equating error”. One may argue that the bias refers to the mean equating error whereas the *RMSD* refers to the variability of equating error. This view is different than the conventional definition of *RMSD* in the literature where it is usually referred to as the total equating error. It is generally expressed in terms of two components: bias due to systematic error and standard error of equating due to sampling error (see Livingston & Kim, 2009, for more details). Despite these conceptual differences it is not surprising that in this study the *av.RMSD* decreased with increasing sample size, test length or the *a*-ratio or base form discrimination as would be predicted in theory. A reduction in the *av.RMSD* of the

equating results based on these factors should not be misconstrued as a revelation or a very important finding because they might be explained by way of sampling theory or other ways of reducing “error” variance. As such the reader should not read too much into the results with respect to the av. RMSD as detailed in Chapters IV and V because they are probably not the most important finding in this study.

In stark contrast, a reduction in the av. BIAS is a more important finding. If, in addition, it is associated with small error variances (small RMSD) this is even more important. This is why the LLIN is the most strongly recommended method from this study. Among the other methods the CE type holds up much better than the PSE type despite being regarded as having “shortcomings” (Kolen & Brennan 2004, pp. 146). The results of this study adds support to the recommendations to practitioners by Sinharay and Holland (2010a) that they should not look down on the methods that belong to the chained equating type “ ... because they appear too simple to be right”, (pp. 282).

Limitations and Directions for Future Research

First, one of the most serious limitations of this study is the lack of symmetry in the levels of factors investigated. Compared to the base form population, none of the alternate form groups are less able. Only one level of test difficulty (SMD=0.25) is higher than the base form difficulty as opposed to three levels (SMD= -0.75, -0.50, -0.10) which are all lower. This lack of balance limits the generalization of the results beyond the conditions examined. For example in this study all the NEAT methods, except for the LLIN, were consistently negatively biased. How would the results have changed if

alternate form groups that were symmetrically less able were included in this study? A similar argument might be made for the level of group differences. Future studies might include broadening the universe of equating conditions to get a more comprehensive picture of the impact of the various conditions.

Second this study used only ten replicates of randomly parallel forms to estimate the accuracy (av.BAIS and av.RMSD) over examinees. How would the results have been different if more replications were used remains an unanswered question?

Third, of all the methods used, only the CARC had its equated scores adjusted to match the identity whenever the observed score on the new form was lower than the pseudo-guessing level. I believe that this might have been to the disadvantage of the other methods where the equated scores were not adjusted.

Fourth, the use of the IRT 3PL model is assumed to fit the data but this by no means guarantees that the findings of this study are generalizable to actual field test data. By the same token, how would the results of this study have changed if the discrimination parameter of the simulated items and the chance level were not fixed?

Fifth, this study did not decide on the number of moments to preserve based on any statistical criteria or fit statistics. The decision was made based on the successful replication of the results reported in Wang (2009). More in depth study might consider finding the optimal degree of smoothing that would introduce the least bias and still give consistently accurate equating results.

Sixth the same study might consider the conditions when the SFRE and SCEE dissociate from one another with respect to the total equating error or when the bias of

the linear and nonlinear methods in the CLIN/SCEE pair or the SFRE/TLIN pair become substantially different from one another. They would very likely behave differently without presmoothing. But then what other conditions, would have made them more interchangeable? Is it large sample size or no difference between groups or what is the minimum test length? These are all legitimate questions that might be addressed in future studies.

Seventh, another important limitation in this study is the definition of the measures of accuracy used in this study. Bias can be positive or negative and the pooling of this measure over examinees to compute the av.BIAS is statistically correct. However, it remains a practical issue because it does not tell where on the score scale the most bias or equating error occurs. Among the most recent reference to this problem includes van der Linden (2010). In his review of the findings by Kane et al., (2009) he underscored the fact that the av.BIAS can be close to zero, whereas the underlying bias function may display bias at nearly every score along the range of raw scores of the new form. This problem has been known for a long time, but the focus of this study was not on the bias at the raw score level rather over the sample of examinees. After all one of the goal of equating is that it should not matter who took which form, the equating results should be interchangeable. Better still they should be interchangeable across various equating methods. To address his problem one could design a similar study as this one that will examine the behavior of the equating transformations at various intervals on the ability range of the raw score scale. A conditional review at different intervals or percentiles

along the raw score scale would shed much needed light on the bias within critical regions of the study.

Ninth, a key component for successful equating under the NEAT design requires that the total test and anchor test correlation be about 0.8 or higher. This study did not focus on this aspect at all and needs to be examined in light of the findings obtained.

Tenth, one further limitation of this study is that in the absence of any validation of the CARC method, the results, claims and comments associated with it may not be valid. The only way to ascertain that it performed or did not perform as intended is to test the algorithm with some already published data set.

REFERENCES

- Albano, A. D. (2010). *equate: Statistical methods for test equating* [Computer software manual]. Available from <http://CRAN.R-project.org/package=equate> (R package)
- Allen, M. J., & Yen, W. M. (2002). *Introduction to Measurement Theory*. Long Grove: Waveland Press, Inc.
- American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.
- Angoff, W. H. (1971). *Scales, norms, and equivalent scores*. In R.L Thorndike (ED.), *Educational Measurement* (2nd ed., pp.508-600). Washington, DC: American Council on Education.
- Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population- independent. *Journal of Educational Measurement*, 23, 327-345.
- Azzalini, A. (2009).sn: The skew-normal and skew-t distributions (version 0.4-11). Available from <http://CRAN.R-project.org/package=sn> (R package)
- Braun, H. I., & Holland, P. W. (1982). *Observed-score test equating: A mathematical analysis of some ETS equating procedures*. In P.W. Holland & D.B Rubin (Eds.), *Test Equating*, (pp.9-49). New York: Academic.

- Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*(3) 225-244
- Divgi, D. R. (1987). *A stable curvilinear alternative to linear equating* (Report CRC 571). Alexandria, VA: Center for Naval Analyses.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.
- Heh, V. (2007). *Equating accuracy using small samples in the random groups design*. (Doctoral dissertation, Ohio University)
http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1178299995
- Hanson, B. A. (1991). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent group design. *Journal of Educational Statistics, 16*, 93-100
- Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsMOOTHING methods in equipercentile equating* (ACT Research Report 94-4). Iowa City, IA: American College Testing, Inc. Iowa City, IA: American College Testing.
- Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Measurement in Education, 10*, 35-43.

- Harris, D. J., & Kolen, M. J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological measurement, 50*, 61-71.
- Harris, D. J. (1993). *Practical issues in equating*. Paper presented at the annual Meeting of the American Educational Research Association, Atlanta, GA.
- Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*, 195-240.
- Holland, P. W., & Dorans, N.J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.187-220). Westport, CT: Preager Publishers.
- Holland, P. W., Dorans, N.J., & Petersen, N.S. (2006). Equating test scores. In C. R. Rao & S. Sinharay (Eds.) *Handbook of statistics, Vol. 26. Psychometrics*. Amsterdam: Elsevier
- Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Technical Report 87-79). Princeton, NJ: Educational Testing Service.
- Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*(2), 133-183.

- Kane, M. T., Mroch, A. A., Suh, Y., & Ripkey, D. R. (2009a). Linear equating for the NEAT design: Parameter substitution models and chained linear relationship models. *Measurement: Interdisciplinary Research and Perspectives*, 7(3), 125–146.
- Kane, M. T., Mroch, A. A., Suh, Y., & Ripkey, D. R. (2009b). Potential Bias in Linear Equating Due to Regression Artifacts. *Measurement: Interdisciplinary Research & Perspective*, 7(3), 123-124
- Kim, S., von Davier, A. A., & Haberman, S. (2006). *An alternative to equating with small samples in the non-equivalent groups anchor test design*. Paper presented at the annual Meeting of the National Council on Measurement in Education, San Francisco, CA.
- Kim, S., von Davier, A. A. & Haberman, S. (2008). Small-Sample Equating Using a Synthetic Linking Function. *Journal of Educational Measurement*, 45: 325–342.
- Kim, S., & Livingston, S.A., (2010), Comparisons among small sample equating methods in a common-item design. *Journal of Educational Measurement*, 47(3), 286-298.
- Kolen M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, 3(1), 97-104.
- Kolen, M. J. (2007). Data Collection Designs and Linking Procedures. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp.31-56). New York: Springer.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

- Livingston, S. A., Dorans, N.J., & Wright, N.K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3, 73-95.
- Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement*, 30, 23-39.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: Educational Testing Service.
- Livingston, S. A., & Kim, S. (2008). *Small-sample equating by the circle-arc method* (Research Report 08-39). Princeton, NJ: Educational Testing Service.
- Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement* 46(3), 330-343
- Livingston, S. A., & Kim, S. (2010), Random-Groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement*, 47,175–185
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8, 452-461.
- Luecht, R. M. (2007). GEN3PLDATA (Version 2). Greensboro, NC.
- Marco, G. L., Petersen, N. S., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D.Weiss (Ed.), *New horizons in testing* (pp. 147-176). New York, NY: Academic Press.

- Mroch, A. A., Suh, Y., Kane, M.T. & Ripkey, D. R. (2009) An evaluation of five linear equating methods for the NEAT design. *Measurement: Interdisciplinary Research & Perspective*, 7(3), 174 —193
- Parshall, C. G., Houghton, P. D. B., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement*,32(1), 37-54.
- Petersen, N. S. (2007). Equating: Best Practices and Challenges to Best Practices. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp.59-72). New York: Springer.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262)
- Puhan, G. (2010). A Comparison of Chained Linear and Poststratification Linear Equating Under Different Testing Conditions. *Journal of Educational Measurement*, 47: 54–75.
- Sinharay, S. & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement*, 44, 249- 275.
- Sinharay, S. & Holland, P. W. (2010a). A new approach to comparing several equating methods in the context of the NEAT design. *Journal of Educational Measurement*, 47(3), 261–285.
- Sinharay, S. & Holland, P. W. (2010b). The Missing Data Assumptions of the NEAT Design and their Implications for Test Equating, *Psychometrika*, 75 (2) 309

- Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement*, 42(4), 309-330.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Suh, Y., Mroch, A. A., Kane, M. T. & Ripkey, D. R. (2009). An empirical comparison of five linear equating methods for the NEAT design. *Measurement: Interdisciplinary Research & Perspective*, 7(3), 147-173
- Team, R. C. (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria,
- van der Linden, W. J. (1997). Book Review [Review of the Test equating: Methods and Practices]. *Psychometrika*, 62(2), 287–290.
- van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika*, 65(4), 437–456.
- van der Linden, W. J. (2006a). Equating error in observed-score equating. *Applied Psychological Measurement*, 30, 355–378.
- van der Linden, W. J. (2006b). Book Review [Review of the Book *The Kernel Method of Test Equating*] *Journal of Educational Measurement*, Vol. 43, No. 3, pp. 291–294
- van der Linden, W. J. (2010). On bias in linear observed-score equating. *Measurement: Interdisciplinary Research and Perspectives*, 8(1), 21–26.
- von Davier, A.A. (2007). Potential Solutions to Practical Equating Issues. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (pp.89-7106). New York: Springer.

- von Davier, A. A. (2008). New results on the linear equating methods for the non-equivalent-groups design. *Journal of Educational and Behavioral Statistics*, 33(2), 186-203.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). *The Kernel Method of Test Equating*. New York: Springer
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). The Chain and Post-Stratification Methods for Observed-Score Equating: Their Relationship to Population Invariance. *Journal of Educational Measurement*, 41, 15–32.
- Wang, T. (2006). *Standard errors of equating for equipercentile equating with log-linear pre-smoothing using the delta method* (CASMA Research Report 14). Center for Advanced Studies in Measurement and Assessment. University of Iowa, Iowa City.
- Wang, T. (2009). Standard errors of equating for the percentile rank–based equipercentile equating with log-linear presmoothing. *Journal of Educational and Behavioral Statistics*, 34(1), pp. 7–23
- Wang, T.; Lee, W., Brennan, R. L.; & Kolen, M. (2006). *A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design*. (CASMA Research Report No. 17). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa

- Wang, T.; Lee, W-C., Brennan, R. L.; & Kolen, M. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item nonequivalent groups design. *Applied Psychological Measurement*, 32, 632–651.
- Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York Springer

APPENDIX A: DESCRIPTIVE STATISTICS FOR AVERAGE BIAS

Table A1: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Values of Test Length for Studies 1-4

Study	Length	Method	Mean	SD	Min	Max	Skew	SE
1	30	CLIN	-0.109	0.228	-0.951	0.473	-0.928	0.013
1	30	SCEE	-0.120	0.231	-0.973	0.465	-0.954	0.013
1	30	SFRE	-0.182	0.247	-0.973	0.433	-0.785	0.014
1	30	TLIN	-0.178	0.248	-0.980	0.430	-0.813	0.014
1	30	LLIN	-0.026	0.234	-0.941	0.654	-0.699	0.013
1	30	CARC	0.050	0.386	-0.816	1.076	0.298	0.022
1	30	IDEN	0.898	1.679	-1.625	4.357	0.468	0.097
2	30	CLIN	-0.157	0.205	-0.822	0.469	-0.556	0.012
2	30	SCEE	-0.176	0.210	-0.866	0.451	-0.604	0.012
2	30	SFRE	-0.215	0.226	-0.907	0.545	-0.556	0.013
2	30	TLIN	-0.205	0.225	-0.893	0.541	-0.572	0.013
2	30	LLIN	-0.102	0.205	-0.787	0.647	-0.338	0.012
2	30	CARC	0.132	0.490	-0.904	1.328	0.464	0.028
2	30	IDEN	1.256	2.006	-1.575	5.105	0.381	0.116
3	60	CLIN	-0.094	0.196	-0.838	0.335	-1.088	0.011
3	60	SCEE	-0.102	0.199	-0.855	0.315	-1.123	0.011
3	60	SFRE	-0.144	0.199	-0.886	0.275	-1.214	0.012
3	60	TLIN	-0.137	0.201	-0.883	0.282	-1.210	0.012
3	60	LLIN	-0.046	0.208	-0.787	0.591	-0.561	0.012
3	60	CARC	0.578	0.387	-0.258	1.708	0.616	0.022
3	60	IDEN	0.932	1.664	-1.540	4.415	0.492	0.096
4	60	CLIN	-0.082	0.175	-0.831	0.513	-1.068	0.010
4	60	SCEE	-0.102	0.184	-0.879	0.534	-1.058	0.011
4	60	SFRE	-0.136	0.189	-0.925	0.392	-1.193	0.011
4	60	TLIN	-0.121	0.184	-0.896	0.378	-1.205	0.011
4	60	LLIN	-0.039	0.176	-0.761	0.659	-0.648	0.010
4	60	CARC	0.708	0.511	-0.293	1.898	0.630	0.030
4	60	IDEN	1.241	2.018	-1.682	5.002	0.355	0.117

Table A2: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Levels of SMD for Study 1 (30 Items, Average Item Discrimination 0.6)

Study	SMD	Method	Mean	SD	Min	Max	Skew	SE
1	0	CLIN	-0.134	0.228	-0.801	0.280	-0.985	0.029
1	0	SCEE	-0.140	0.234	-0.824	0.282	-0.995	0.030
1	0	SFRE	-0.197	0.245	-0.904	0.232	-0.858	0.032
1	0	TLIN	-0.197	0.245	-0.906	0.232	-0.869	0.032
1	0	LLIN	-0.057	0.240	-0.700	0.363	-0.693	0.031
1	0	CARC	-0.123	0.221	-0.782	0.249	-1.132	0.029
1	0	IDEN	-0.072	0.212	-0.703	0.280	-1.075	0.027
1	-0.1	CLIN	-0.099	0.235	-0.855	0.407	-1.114	0.030
1	-0.1	SCEE	-0.107	0.239	-0.881	0.412	-1.126	0.031
1	-0.1	SFRE	-0.178	0.262	-0.973	0.241	-1.034	0.034
1	-0.1	TLIN	-0.177	0.262	-0.980	0.239	-1.060	0.034
1	-0.1	LLIN	-0.004	0.238	-0.700	0.613	-0.771	0.031
1	-0.1	CARC	-0.015	0.243	-0.816	0.450	-1.310	0.031
1	-0.1	IDEN	0.399	0.285	-0.420	0.843	-1.004	0.037
1	-0.5	CLIN	-0.120	0.263	-0.951	0.473	-0.709	0.034
1	-0.5	SCEE	-0.134	0.265	-0.973	0.465	-0.726	0.034
1	-0.5	SFRE	-0.192	0.278	-0.962	0.433	-0.358	0.036
1	-0.5	TLIN	-0.185	0.278	-0.960	0.430	-0.395	0.036
1	-0.5	LLIN	-0.041	0.268	-0.941	0.525	-1.003	0.035
1	-0.5	CARC	0.249	0.358	-0.794	0.955	-0.764	0.046
1	-0.5	IDEN	2.137	0.710	0.856	2.997	-0.454	0.092
1	-0.75	CLIN	-0.093	0.186	-0.735	0.215	-1.120	0.024
1	-0.75	SCEE	-0.114	0.187	-0.759	0.165	-1.130	0.024
1	-0.75	SFRE	-0.173	0.209	-0.847	0.168	-0.890	0.027
1	-0.75	TLIN	-0.159	0.213	-0.847	0.214	-0.937	0.027
1	-0.75	LLIN	-0.012	0.186	-0.599	0.340	-0.721	0.024
1	-0.75	CARC	0.427	0.378	-0.522	1.076	-0.422	0.049
1	-0.75	IDEN	3.218	0.949	1.579	4.357	-0.360	0.123
1	0.25	CLIN	-0.101	0.225	-0.731	0.447	-0.638	0.029
1	0.25	SCEE	-0.106	0.229	-0.762	0.418	-0.733	0.030
1	0.25	SFRE	-0.171	0.243	-0.795	0.265	-0.784	0.031
1	0.25	TLIN	-0.171	0.245	-0.802	0.279	-0.789	0.032
1	0.25	LLIN	-0.014	0.234	-0.643	0.654	0.066	0.030
1	0.25	CARC	-0.287	0.203	-0.779	0.256	0.101	0.026
1	0.25	IDEN	-1.191	0.262	-1.625	-0.568	0.622	0.034

Table A3: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Levels of SMD for Study 2 (30 Items, Average Item Discrimination 1.0)

Study	SMD	Method	Mean	SD	Min	Max	Skew	SE
2	0	CLIN	-0.148	0.190	-0.805	0.132	-1.201	0.025
2	0	SCEE	-0.158	0.197	-0.843	0.120	-1.224	0.025
2	0	SFRE	-0.199	0.218	-0.876	0.112	-1.110	0.028
2	0	TLIN	-0.198	0.218	-0.866	0.106	-1.106	0.028
2	0	LLIN	-0.091	0.184	-0.735	0.173	-0.988	0.024
2	0	CARC	-0.114	0.180	-0.757	0.179	-1.369	0.023
2	0	IDEN	0.012	0.147	-0.476	0.290	-1.100	0.019
2	-0.1	CLIN	-0.137	0.183	-0.646	0.200	-0.513	0.024
2	-0.1	SCEE	-0.148	0.193	-0.683	0.189	-0.592	0.025
2	-0.1	SFRE	-0.192	0.213	-0.778	0.213	-0.505	0.027
2	-0.1	TLIN	-0.190	0.209	-0.762	0.219	-0.464	0.027
2	-0.1	LLIN	-0.076	0.182	-0.525	0.384	-0.288	0.024
2	-0.1	CARC	0.033	0.209	-0.547	0.456	-0.698	0.027
2	-0.1	IDEN	0.642	0.230	0.113	1.237	-0.349	0.030
2	-0.5	CLIN	-0.171	0.198	-0.822	0.233	-0.720	0.026
2	-0.5	SCEE	-0.194	0.202	-0.866	0.207	-0.784	0.026
2	-0.5	SFRE	-0.227	0.226	-0.907	0.225	-0.614	0.029
2	-0.5	TLIN	-0.212	0.228	-0.893	0.250	-0.597	0.029
2	-0.5	LLIN	-0.123	0.182	-0.741	0.274	-0.660	0.023
2	-0.5	CARC	0.464	0.354	-0.514	1.005	-0.747	0.046
2	-0.5	IDEN	2.778	0.584	1.713	3.507	-0.534	0.075
2	-0.75	CLIN	-0.181	0.200	-0.758	0.152	-0.671	0.026
2	-0.75	SCEE	-0.223	0.199	-0.797	0.107	-0.707	0.026
2	-0.75	SFRE	-0.258	0.210	-0.844	0.113	-0.665	0.027
2	-0.75	TLIN	-0.226	0.215	-0.827	0.125	-0.694	0.028
2	-0.75	LLIN	-0.130	0.202	-0.678	0.217	-0.555	0.026
2	-0.75	CARC	0.677	0.443	-0.307	1.328	-0.485	0.057
2	-0.75	IDEN	4.130	0.830	2.669	5.105	-0.496	0.107
2	0.25	CLIN	-0.147	0.250	-0.673	0.469	-0.107	0.032
2	0.25	SCEE	-0.157	0.252	-0.656	0.451	-0.209	0.033
2	0.25	SFRE	-0.200	0.258	-0.742	0.545	-0.182	0.033
2	0.25	TLIN	-0.199	0.257	-0.738	0.541	-0.176	0.033
2	0.25	LLIN	-0.088	0.265	-0.787	0.647	0.055	0.034
2	0.25	CARC	-0.398	0.222	-0.904	0.284	0.558	0.029
2	0.25	IDEN	-1.282	0.163	-1.575	-0.888	0.595	0.021

Table A4: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Levels of SMD for Study 3 (60 Items, Average Item Discrimination 0.6)

Study	SMD	Method	Mean	SD	Min	Max	Skew	SE
3	0	CLIN	-0.124	0.181	-0.786	0.126	-1.428	0.023
3	0	SCEE	-0.129	0.186	-0.807	0.129	-1.412	0.024
3	0	SFRE	-0.168	0.185	-0.857	0.076	-1.567	0.024
3	0	TLIN	-0.165	0.185	-0.854	0.085	-1.560	0.024
3	0	LLIN	-0.078	0.190	-0.709	0.295	-0.811	0.025
3	0	CARC	0.389	0.180	-0.258	0.652	-1.392	0.023
3	0	IDEN	-0.031	0.174	-0.442	0.371	-0.445	0.022
3	-0.1	CLIN	-0.085	0.186	-0.639	0.283	-1.001	0.024
3	-0.1	SCEE	-0.090	0.189	-0.647	0.285	-0.992	0.024
3	-0.1	SFRE	-0.135	0.197	-0.758	0.130	-1.223	0.025
3	-0.1	TLIN	-0.131	0.199	-0.764	0.130	-1.231	0.026
3	-0.1	LLIN	-0.034	0.189	-0.498	0.503	-0.229	0.024
3	-0.1	CARC	0.493	0.206	-0.107	0.909	-0.959	0.027
3	-0.1	IDEN	0.403	0.266	-0.285	0.910	-0.678	0.034
3	-0.5	CLIN	-0.076	0.210	-0.838	0.310	-1.174	0.027
3	-0.5	SCEE	-0.086	0.212	-0.855	0.286	-1.205	0.027
3	-0.5	SFRE	-0.133	0.209	-0.886	0.214	-1.319	0.027
3	-0.5	TLIN	-0.124	0.212	-0.883	0.221	-1.318	0.027
3	-0.5	LLIN	-0.023	0.224	-0.787	0.426	-0.648	0.029
3	-0.5	CARC	0.798	0.353	-0.201	1.359	-0.664	0.046
3	-0.5	IDEN	2.179	0.706	0.824	3.323	-0.421	0.091
3	-0.75	CLIN	-0.092	0.218	-0.662	0.335	-0.961	0.028
3	-0.75	SCEE	-0.111	0.217	-0.680	0.315	-0.983	0.028
3	-0.75	SFRE	-0.150	0.215	-0.746	0.275	-1.005	0.028
3	-0.75	TLIN	-0.133	0.219	-0.745	0.282	-1.037	0.028
3	-0.75	LLIN	-0.047	0.238	-0.595	0.591	-0.480	0.031
3	-0.75	CARC	0.964	0.422	0.032	1.708	-0.373	0.054
3	-0.75	IDEN	3.231	0.948	1.593	4.415	-0.381	0.122
3	0.25	CLIN	-0.093	0.187	-0.634	0.219	-0.941	0.024
3	0.25	SCEE	-0.096	0.190	-0.651	0.198	-1.003	0.025
3	0.25	SFRE	-0.134	0.193	-0.683	0.176	-0.966	0.025
3	0.25	TLIN	-0.131	0.193	-0.680	0.186	-0.956	0.025
3	0.25	LLIN	-0.050	0.198	-0.582	0.297	-0.678	0.026
3	0.25	CARC	0.249	0.165	-0.173	0.668	0.012	0.021
3	0.25	IDEN	-1.123	0.227	-1.540	-0.619	0.630	0.029

Table A5: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Levels of SMD for Study 4 (60 Items, Average Item Discrimination 1.0)

Study	SMD	Method	Mean	SD	Min	Max	Skew	SE
4	0	CLIN	-0.078	0.207	-0.831	0.290	-1.584	0.027
4	0	SCEE	-0.088	0.218	-0.879	0.305	-1.636	0.028
4	0	SFRE	-0.122	0.224	-0.925	0.262	-1.647	0.029
4	0	TLIN	-0.117	0.218	-0.896	0.254	-1.595	0.028
4	0	LLIN	-0.035	0.206	-0.761	0.428	-1.243	0.027
4	0	CARC	0.441	0.198	-0.293	0.795	-1.733	0.026
4	0	IDEN	0.040	0.151	-0.550	0.323	-1.694	0.020
4	-0.1	CLIN	-0.091	0.181	-0.707	0.189	-1.559	0.023
4	-0.1	SCEE	-0.100	0.196	-0.771	0.195	-1.518	0.025
4	-0.1	SFRE	-0.134	0.204	-0.824	0.142	-1.527	0.026
4	-0.1	TLIN	-0.130	0.194	-0.767	0.130	-1.516	0.025
4	-0.1	LLIN	-0.048	0.178	-0.641	0.254	-1.258	0.023
4	-0.1	CARC	0.550	0.209	-0.113	0.859	-1.440	0.027
4	-0.1	IDEN	0.580	0.239	-0.033	0.903	-1.069	0.031
4	-0.5	CLIN	-0.105	0.167	-0.751	0.292	-1.054	0.022
4	-0.5	SCEE	-0.127	0.174	-0.787	0.238	-1.102	0.022
4	-0.5	SFRE	-0.162	0.185	-0.866	0.257	-1.097	0.024
4	-0.5	TLIN	-0.145	0.183	-0.852	0.297	-1.079	0.024
4	-0.5	LLIN	-0.062	0.161	-0.641	0.286	-0.796	0.021
4	-0.5	CARC	1.029	0.340	0.079	1.487	-0.661	0.044
4	-0.5	IDEN	2.795	0.593	1.467	3.461	-0.575	0.077
4	-0.75	CLIN	-0.072	0.155	-0.479	0.344	-0.418	0.020
4	-0.75	SCEE	-0.120	0.155	-0.529	0.313	-0.345	0.020
4	-0.75	SFRE	-0.154	0.157	-0.596	0.229	-0.559	0.020
4	-0.75	TLIN	-0.112	0.161	-0.566	0.251	-0.692	0.021
4	-0.75	LLIN	-0.029	0.161	-0.385	0.446	-0.011	0.021
4	-0.75	CARC	1.350	0.438	0.488	1.898	-0.474	0.057
4	-0.75	IDEN	4.123	0.785	2.745	5.002	-0.523	0.101
4	0.25	CLIN	-0.064	0.162	-0.465	0.513	0.352	0.021
4	0.25	SCEE	-0.075	0.174	-0.498	0.534	0.229	0.022
4	0.25	SFRE	-0.110	0.167	-0.559	0.392	-0.063	0.022
4	0.25	TLIN	-0.102	0.161	-0.545	0.378	0.020	0.021
4	0.25	LLIN	-0.022	0.174	-0.379	0.659	0.630	0.022
4	0.25	CARC	0.168	0.133	-0.084	0.521	0.697	0.017
4	0.25	IDEN	-1.334	0.200	-1.682	-0.921	0.593	0.026

Table A6: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Levels of Average Item Discrimination for Study 1 (30 Items, Average Item Discrimination 0.6)

Study	Discrimination	Method	Average	SD	Min	Max	Skew	SE
1	0.3	CLIN	-0.229	0.284	-0.951	0.382	-0.453	0.028
1	0.3	SCEE	-0.246	0.287	-0.973	0.359	-0.465	0.029
1	0.3	SFRE	-0.276	0.313	-0.973	0.407	-0.461	0.031
1	0.3	TLIN	-0.278	0.313	-0.980	0.406	-0.462	0.031
1	0.3	LLIN	-0.170	0.258	-0.941	0.457	-0.475	0.026
1	0.3	CARC	-0.173	0.296	-0.816	0.483	-0.348	0.030
1	0.3	IDEN	0.415	1.075	-1.427	2.305	0.231	0.107
1	0.6	CLIN	-0.072	0.168	-0.634	0.447	0.054	0.017
1	0.6	SCEE	-0.083	0.169	-0.633	0.418	0.028	0.017
1	0.6	SFRE	-0.153	0.201	-0.739	0.265	-0.383	0.020
1	0.6	TLIN	-0.150	0.201	-0.726	0.279	-0.376	0.020
1	0.6	LLIN	0.023	0.166	-0.522	0.654	0.608	0.017
1	0.6	CARC	0.090	0.294	-0.746	0.635	-0.214	0.029
1	0.6	IDEN	0.996	1.686	-1.424	3.699	0.188	0.169
1	0.9	CLIN	-0.027	0.158	-0.591	0.473	-0.562	0.016
1	0.9	SCEE	-0.032	0.158	-0.600	0.465	-0.580	0.016
1	0.9	SFRE	-0.118	0.179	-0.599	0.433	-0.039	0.018
1	0.9	TLIN	-0.105	0.177	-0.576	0.430	-0.047	0.018
1	0.9	LLIN	0.070	0.195	-0.635	0.525	-0.762	0.019
1	0.9	CARC	0.234	0.436	-0.504	1.076	0.186	0.044
1	0.9	IDEN	1.284	2.029	-1.625	4.357	0.195	0.203

Table A7: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Levels of Average Item Discrimination for Study 2 (30 Items, Average Item Discrimination 1.0)

Study	Discrimination	Method	Average	SD	Min	Max	Skew	SE
2	0.5	CLIN	-0.257	0.246	-0.822	0.469	-0.059	0.025
2	0.5	SCEE	-0.292	0.247	-0.866	0.423	-0.073	0.025
2	0.5	SFRE	-0.309	0.265	-0.907	0.297	-0.275	0.026
2	0.5	TLIN	-0.293	0.267	-0.893	0.314	-0.292	0.027
2	0.5	LLIN	-0.216	0.235	-0.741	0.647	0.343	0.023
2	0.5	CARC	-0.110	0.296	-0.757	0.568	-0.132	0.030
2	0.5	IDEN	0.840	1.508	-1.417	3.334	0.206	0.151
2	1	CLIN	-0.120	0.174	-0.536	0.468	0.286	0.017
2	1	SCEE	-0.134	0.175	-0.579	0.451	0.225	0.018
2	1	SFRE	-0.182	0.207	-0.673	0.545	0.045	0.021
2	1	TLIN	-0.175	0.207	-0.677	0.541	0.012	0.021
2	1	LLIN	-0.057	0.161	-0.474	0.495	0.629	0.016
2	1	CARC	0.195	0.448	-0.715	1.103	0.022	0.045
2	1	IDEN	1.366	2.073	-1.447	4.659	0.247	0.207
2	1.5	CLIN	-0.093	0.143	-0.673	0.220	-1.174	0.014
2	1.5	SCEE	-0.103	0.145	-0.645	0.207	-1.104	0.014
2	1.5	SFRE	-0.155	0.165	-0.691	0.225	-0.773	0.017
2	1.5	TLIN	-0.147	0.164	-0.728	0.250	-0.800	0.016
2	1.5	LLIN	-0.032	0.161	-0.787	0.384	-1.168	0.016
2	1.5	CARC	0.313	0.584	-0.904	1.328	0.086	0.058
2	1.5	IDEN	1.561	2.302	-1.575	5.105	0.241	0.230

Table A8: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Levels of Average Item Discrimination for Study 3 (60 Items, Average Item Discrimination 0.6)

Study	Discrimination	Method	Average	SD	Min	Max	Skew	SE
3	0.3	CLIN	-0.210	0.249	-0.838	0.335	-0.294	0.025
3	0.3	SCEE	-0.226	0.248	-0.855	0.315	-0.307	0.025
3	0.3	SFRE	-0.242	0.266	-0.886	0.275	-0.422	0.027
3	0.3	TLIN	-0.238	0.267	-0.883	0.282	-0.424	0.027
3	0.3	LLIN	-0.178	0.232	-0.787	0.411	-0.100	0.023
3	0.3	CARC	0.347	0.260	-0.258	1.002	-0.136	0.026
3	0.3	IDEN	0.458	1.059	-1.297	2.288	0.248	0.106
3	0.6	CLIN	-0.087	0.127	-0.545	0.208	-0.917	0.013
3	0.6	SCEE	-0.091	0.128	-0.560	0.206	-0.966	0.013
3	0.6	SFRE	-0.133	0.149	-0.524	0.180	-0.585	0.015
3	0.6	TLIN	-0.129	0.150	-0.516	0.194	-0.546	0.015
3	0.6	LLIN	-0.039	0.125	-0.595	0.224	-1.231	0.012
3	0.6	CARC	0.592	0.296	-0.037	1.151	0.292	0.030
3	0.6	IDEN	1.015	1.677	-1.270	3.562	0.224	0.168
3	0.9	CLIN	0.013	0.115	-0.539	0.285	-0.967	0.012
3	0.9	SCEE	0.010	0.115	-0.563	0.285	-1.153	0.012
3	0.9	SFRE	-0.055	0.096	-0.533	0.170	-1.214	0.010
3	0.9	TLIN	-0.043	0.095	-0.510	0.221	-1.084	0.010
3	0.9	LLIN	0.077	0.169	-0.572	0.591	-0.201	0.017
3	0.9	CARC	0.796	0.442	0.018	1.708	0.240	0.044
3	0.9	IDEN	1.322	2.010	-1.540	4.415	0.208	0.201

Table A9: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Levels of Average Item Discrimination for Study 4 (60 Items, Average Item Discrimination 1.0)

Study	Discrimination	Method	Average	SD	Min	Max	Skew	SE
4	0.5	CLIN	-0.186	0.218	-0.831	0.292	-0.698	0.022
4	0.5	SCEE	-0.233	0.216	-0.879	0.238	-0.703	0.022
4	0.5	SFRE	-0.252	0.234	-0.925	0.257	-0.733	0.023
4	0.5	TLIN	-0.217	0.239	-0.896	0.297	-0.736	0.024
4	0.5	LLIN	-0.153	0.199	-0.761	0.286	-0.626	0.020
4	0.5	CARC	0.459	0.300	-0.293	1.071	-0.174	0.030
4	0.5	IDEN	0.845	1.513	-1.382	3.281	0.257	0.151
4	1	CLIN	-0.050	0.127	-0.378	0.344	0.023	0.013
4	1	SCEE	-0.060	0.126	-0.406	0.313	-0.089	0.013
4	1	SFRE	-0.098	0.137	-0.457	0.229	-0.185	0.014
4	1	TLIN	-0.092	0.139	-0.440	0.251	-0.138	0.014
4	1	LLIN	-0.004	0.130	-0.375	0.446	0.034	0.013
4	1	CARC	0.768	0.479	0.011	1.798	0.340	0.048
4	1	IDEN	1.344	2.101	-1.516	4.544	0.220	0.210
4	1.5	CLIN	-0.010	0.106	-0.233	0.513	1.377	0.011
4	1.5	SCEE	-0.013	0.113	-0.285	0.534	1.145	0.011
4	1.5	SFRE	-0.058	0.112	-0.337	0.392	0.521	0.011
4	1.5	TLIN	-0.055	0.105	-0.338	0.378	0.677	0.011
4	1.5	LLIN	0.039	0.129	-0.228	0.659	1.431	0.013
4	1.5	CARC	0.896	0.606	-0.072	1.898	0.214	0.061
4	1.5	IDEN	1.533	2.312	-1.682	5.002	0.207	0.231

Table A10: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for Various Sample Sizes for Study 1 (30 Items, Average Item Discrimination 0.6)

Study	Size	Method	Mean	SD	Min	Max	Skew	SE
1	25	CLIN	-0.112	0.289	-0.731	0.473	-0.222	0.037
1	25	SCEE	-0.123	0.290	-0.762	0.465	-0.259	0.037
1	25	SFRE	-0.183	0.290	-0.824	0.433	-0.263	0.037
1	25	TLIN	-0.177	0.292	-0.833	0.430	-0.285	0.038
1	25	LLIN	-0.031	0.315	-0.700	0.654	-0.092	0.041
1	25	CARC	0.044	0.430	-0.779	1.076	0.374	0.055
1	25	IDEN	0.904	1.707	-1.625	4.357	0.461	0.220
1	50	CLIN	-0.138	0.237	-0.855	0.382	-0.646	0.031
1	50	SCEE	-0.148	0.242	-0.881	0.359	-0.660	0.031
1	50	SFRE	-0.208	0.263	-0.973	0.407	-0.467	0.034
1	50	TLIN	-0.203	0.263	-0.980	0.406	-0.529	0.034
1	50	LLIN	-0.058	0.241	-0.716	0.451	-0.578	0.031
1	50	CARC	0.028	0.361	-0.816	0.870	0.037	0.047
1	50	IDEN	0.897	1.666	-1.587	4.260	0.441	0.215
1	100	CLIN	-0.110	0.232	-0.951	0.247	-1.558	0.030
1	100	SCEE	-0.121	0.236	-0.973	0.238	-1.544	0.030
1	100	SFRE	-0.184	0.252	-0.962	0.200	-1.224	0.032
1	100	TLIN	-0.179	0.254	-0.960	0.198	-1.230	0.033
1	100	LLIN	-0.025	0.235	-0.941	0.361	-1.367	0.030
1	100	CARC	0.047	0.399	-0.794	0.967	0.157	0.052
1	100	IDEN	0.884	1.695	-1.451	4.303	0.474	0.219
1	200	CLIN	-0.089	0.185	-0.695	0.200	-1.328	0.024
1	200	SCEE	-0.101	0.189	-0.713	0.191	-1.338	0.024
1	200	SFRE	-0.166	0.211	-0.788	0.111	-1.099	0.027
1	200	TLIN	-0.162	0.213	-0.789	0.133	-1.116	0.027
1	200	LLIN	0.000	0.182	-0.580	0.395	-0.851	0.023
1	200	CARC	0.069	0.380	-0.677	1.005	0.552	0.049
1	200	IDEN	0.899	1.695	-1.520	4.306	0.462	0.219
1	400	CLIN	-0.099	0.183	-0.653	0.183	-1.753	0.024
1	400	SCEE	-0.109	0.187	-0.675	0.165	-1.754	0.024
1	400	SFRE	-0.171	0.217	-0.758	0.169	-1.222	0.028
1	400	TLIN	-0.167	0.217	-0.761	0.170	-1.257	0.028
1	400	LLIN	-0.015	0.173	-0.529	0.251	-1.424	0.022
1	400	CARC	0.063	0.368	-0.642	0.848	0.289	0.047
1	400	IDEN	0.908	1.686	-1.526	4.273	0.453	0.218

Table A11: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for Various Sample Sizes for Study 2 (30 Items, Average Item Discrimination 1.0)

Study	Size	Method	Mean	SD	Min	Max	Skew	SE
2	25	CLIN	-0.117	0.269	-0.673	0.469	-0.065	0.035
2	25	SCEE	-0.140	0.270	-0.645	0.451	-0.065	0.035
2	25	SFRE	-0.184	0.273	-0.691	0.545	-0.014	0.035
2	25	TLIN	-0.171	0.274	-0.728	0.541	-0.082	0.035
2	25	LLIN	-0.055	0.293	-0.787	0.647	-0.086	0.038
2	25	CARC	0.164	0.495	-0.904	1.211	0.178	0.064
2	25	IDEN	1.273	2.018	-1.561	5.105	0.390	0.260
2	50	CLIN	-0.197	0.218	-0.822	0.220	-0.642	0.028
2	50	SCEE	-0.218	0.225	-0.866	0.207	-0.640	0.029
2	50	SFRE	-0.257	0.243	-0.907	0.225	-0.521	0.031
2	50	TLIN	-0.246	0.241	-0.893	0.250	-0.534	0.031
2	50	LLIN	-0.142	0.211	-0.741	0.212	-0.563	0.027
2	50	CARC	0.095	0.508	-0.715	1.305	0.568	0.066
2	50	IDEN	1.233	2.027	-1.575	5.095	0.357	0.262
2	100	CLIN	-0.165	0.196	-0.805	0.137	-1.240	0.025
2	100	SCEE	-0.184	0.204	-0.843	0.107	-1.229	0.026
2	100	SFRE	-0.222	0.221	-0.876	0.113	-1.031	0.029
2	100	TLIN	-0.211	0.219	-0.866	0.125	-1.056	0.028
2	100	LLIN	-0.112	0.187	-0.735	0.151	-1.209	0.024
2	100	CARC	0.125	0.488	-0.757	1.118	0.353	0.063
2	100	IDEN	1.255	2.010	-1.423	5.004	0.370	0.259
2	200	CLIN	-0.148	0.152	-0.550	0.126	-0.749	0.020
2	200	SCEE	-0.166	0.160	-0.592	0.120	-0.784	0.021
2	200	SFRE	-0.203	0.186	-0.686	0.097	-0.717	0.024
2	200	TLIN	-0.194	0.185	-0.671	0.106	-0.710	0.024
2	200	LLIN	-0.095	0.141	-0.412	0.150	-0.386	0.018
2	200	CARC	0.143	0.480	-0.667	1.265	0.610	0.062
2	200	IDEN	1.261	2.026	-1.447	5.021	0.372	0.262
2	400	CLIN	-0.158	0.167	-0.647	0.124	-1.404	0.022
2	400	SCEE	-0.174	0.174	-0.687	0.094	-1.425	0.023
2	400	SFRE	-0.212	0.196	-0.758	0.112	-1.193	0.025
2	400	TLIN	-0.205	0.196	-0.746	0.132	-1.173	0.025
2	400	LLIN	-0.104	0.155	-0.535	0.118	-1.008	0.020
2	400	CARC	0.135	0.493	-0.721	1.328	0.589	0.064
2	400	IDEN	1.258	2.017	-1.457	5.004	0.377	0.260

Table A12: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Sample Sizes for Study 3 (60 Items, Average Item Discrimination 0.6)

Study	Size	Method	Mean	SD	Min	Max	Skew	SE
3	25	CLIN	-0.093	0.256	-0.786	0.335	-0.494	0.033
3	25	SCEE	-0.103	0.257	-0.807	0.315	-0.536	0.033
3	25	SFRE	-0.146	0.248	-0.857	0.275	-0.613	0.032
3	25	TLIN	-0.138	0.250	-0.854	0.282	-0.611	0.032
3	25	LLIN	-0.043	0.280	-0.709	0.591	-0.177	0.036
3	25	CARC	0.572	0.403	-0.258	1.708	0.460	0.052
3	25	IDEN	0.941	1.688	-1.427	4.326	0.450	0.218
3	50	CLIN	-0.084	0.211	-0.838	0.255	-1.406	0.027
3	50	SCEE	-0.092	0.212	-0.855	0.245	-1.453	0.027
3	50	SFRE	-0.132	0.216	-0.886	0.203	-1.446	0.028
3	50	TLIN	-0.125	0.219	-0.883	0.209	-1.444	0.028
3	50	LLIN	-0.037	0.219	-0.787	0.426	-0.991	0.028
3	50	CARC	0.589	0.402	-0.201	1.541	0.472	0.052
3	50	IDEN	0.936	1.666	-1.503	4.415	0.509	0.215
3	100	CLIN	-0.092	0.177	-0.640	0.210	-1.348	0.023
3	100	SCEE	-0.099	0.180	-0.657	0.208	-1.349	0.023
3	100	SFRE	-0.140	0.181	-0.716	0.169	-1.566	0.023
3	100	TLIN	-0.133	0.183	-0.714	0.174	-1.577	0.024
3	100	LLIN	-0.045	0.190	-0.582	0.403	-0.524	0.025
3	100	CARC	0.582	0.380	-0.173	1.519	0.644	0.049
3	100	IDEN	0.931	1.669	-1.540	4.309	0.481	0.216
3	200	CLIN	-0.103	0.167	-0.597	0.126	-1.348	0.022
3	200	SCEE	-0.111	0.171	-0.614	0.129	-1.325	0.022
3	200	SFRE	-0.153	0.177	-0.673	0.095	-1.316	0.023
3	200	TLIN	-0.147	0.179	-0.671	0.106	-1.323	0.023
3	200	LLIN	-0.054	0.169	-0.514	0.295	-0.874	0.022
3	200	CARC	0.573	0.376	-0.147	1.481	0.689	0.049
3	200	IDEN	0.924	1.675	-1.419	4.251	0.477	0.216
3	400	CLIN	-0.100	0.161	-0.607	0.112	-1.580	0.021
3	400	SCEE	-0.107	0.164	-0.623	0.086	-1.579	0.021
3	400	SFRE	-0.146	0.169	-0.665	0.060	-1.610	0.022
3	400	TLIN	-0.140	0.171	-0.663	0.066	-1.585	0.022
3	400	LLIN	-0.053	0.166	-0.545	0.286	-0.907	0.021
3	400	CARC	0.576	0.387	-0.003	1.549	0.790	0.050
3	400	IDEN	0.925	1.676	-1.396	4.314	0.494	0.216

Table A13: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for the Various Sample Sizes for Study 4 (60 Items, Average Item Discrimination 1.0)

Study	Size	Method	Mean	SD	Min	Max	Skew	SE
4	25	CLIN	-0.100	0.216	-0.831	0.513	-0.466	0.028
4	25	SCEE	-0.123	0.225	-0.879	0.534	-0.467	0.029
4	25	SFRE	-0.156	0.220	-0.925	0.392	-0.771	0.028
4	25	TLIN	-0.136	0.214	-0.896	0.378	-0.797	0.028
4	25	LLIN	-0.062	0.231	-0.761	0.659	-0.023	0.030
4	25	CARC	0.684	0.504	-0.293	1.892	0.540	0.065
4	25	IDEN	1.215	2.023	-1.561	4.947	0.361	0.261
4	50	CLIN	-0.077	0.223	-0.751	0.344	-0.907	0.029
4	50	SCEE	-0.098	0.233	-0.797	0.313	-0.924	0.030
4	50	SFRE	-0.131	0.236	-0.866	0.257	-1.059	0.030
4	50	TLIN	-0.116	0.232	-0.852	0.297	-1.067	0.030
4	50	LLIN	-0.035	0.222	-0.674	0.446	-0.600	0.029
4	50	CARC	0.709	0.522	-0.208	1.898	0.608	0.067
4	50	IDEN	1.244	2.028	-1.682	5.002	0.342	0.262
4	100	CLIN	-0.092	0.138	-0.530	0.134	-1.319	0.018
4	100	SCEE	-0.113	0.150	-0.576	0.123	-1.231	0.019
4	100	SFRE	-0.148	0.157	-0.619	0.092	-1.141	0.020
4	100	TLIN	-0.133	0.152	-0.608	0.102	-1.163	0.020
4	100	LLIN	-0.048	0.137	-0.473	0.169	-1.057	0.018
4	100	CARC	0.702	0.514	0.007	1.811	0.623	0.066
4	100	IDEN	1.250	2.039	-1.544	4.969	0.345	0.263
4	200	CLIN	-0.084	0.138	-0.590	0.090	-1.694	0.018
4	200	SCEE	-0.103	0.149	-0.624	0.086	-1.528	0.019
4	200	SFRE	-0.137	0.161	-0.687	0.073	-1.446	0.021
4	200	TLIN	-0.124	0.157	-0.671	0.080	-1.527	0.020
4	200	LLIN	-0.041	0.128	-0.501	0.185	-1.351	0.017
4	200	CARC	0.709	0.512	-0.008	1.808	0.629	0.066
4	200	IDEN	1.245	2.041	-1.569	4.918	0.346	0.263
4	400	CLIN	-0.055	0.138	-0.508	0.135	-1.802	0.018
4	400	SCEE	-0.073	0.147	-0.543	0.123	-1.766	0.019
4	400	SFRE	-0.110	0.155	-0.597	0.096	-1.720	0.020
4	400	TLIN	-0.098	0.153	-0.581	0.125	-1.679	0.020
4	400	LLIN	-0.010	0.132	-0.430	0.240	-1.378	0.017
4	400	CARC	0.735	0.520	-0.084	1.877	0.674	0.067
4	400	IDEN	1.249	2.029	-1.564	4.926	0.344	0.262

Table A14: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for Various Levels of STD for Study 1 (30 Items, Average Item Discrimination 0.6)

Study	STD	Method	Mean	SD	Min	Max	Range	Skew	SE
1	0	CLIN	0.034	0.141	-0.262	0.447	0.709	0.471	0.016
1	0	SCEE	0.026	0.140	-0.284	0.418	0.702	0.387	0.016
1	0	SFRE	0.021	0.131	-0.299	0.407	0.706	-0.003	0.015
1	0	TLIN	0.026	0.130	-0.273	0.406	0.679	0.043	0.015
1	0	LLIN	0.044	0.167	-0.247	0.654	0.901	1.062	0.019
1	0	CARC	0.184	0.338	-0.504	1.076	1.580	0.548	0.039
1	0	IDEN	0.961	1.679	-1.587	4.306	5.893	0.420	0.194
1	0.05	CLIN	-0.027	0.154	-0.591	0.473	1.063	-0.101	0.018
1	0.05	SCEE	-0.036	0.154	-0.603	0.465	1.068	-0.104	0.018
1	0.05	SFRE	-0.067	0.139	-0.501	0.433	0.934	0.214	0.016
1	0.05	TLIN	-0.063	0.138	-0.502	0.430	0.932	0.162	0.016
1	0.05	LLIN	0.018	0.183	-0.700	0.525	1.225	-0.399	0.021
1	0.05	CARC	0.124	0.339	-0.577	0.955	1.532	0.598	0.039
1	0.05	IDEN	0.925	1.686	-1.625	4.340	5.965	0.438	0.195
1	0.1	CLIN	-0.150	0.168	-0.704	0.295	0.999	-0.767	0.019
1	0.1	SCEE	-0.161	0.169	-0.724	0.288	1.011	-0.735	0.020
1	0.1	SFRE	-0.220	0.145	-0.696	0.207	0.903	-0.639	0.017
1	0.1	TLIN	-0.215	0.147	-0.694	0.208	0.902	-0.616	0.017
1	0.1	LLIN	-0.071	0.200	-0.716	0.403	1.119	-0.831	0.023
1	0.1	CARC	0.012	0.347	-0.557	0.870	1.428	0.704	0.040
1	0.1	IDEN	0.883	1.683	-1.532	4.303	5.835	0.472	0.194
1	0.25	CLIN	-0.295	0.272	-0.951	0.146	1.097	-0.430	0.031
1	0.25	SCEE	-0.310	0.277	-0.973	0.144	1.117	-0.419	0.032
1	0.25	SFRE	-0.464	0.228	-0.973	-0.105	0.867	-0.397	0.026
1	0.25	TLIN	-0.459	0.233	-0.980	-0.080	0.900	-0.398	0.027
1	0.25	LLIN	-0.093	0.325	-0.941	0.451	1.391	-0.436	0.038
1	0.25	CARC	-0.118	0.445	-0.816	0.912	1.728	0.440	0.051
1	0.25	IDEN	0.825	1.698	-1.427	4.357	5.784	0.518	0.196

Table A15: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for Various Levels of STD for Study 2 (30 Items, Average Item Discrimination 1.0)

Study	STD	Method	Mean	SD	Min	Max	Range	Skew	SE
2	0	CLIN	-0.050	0.150	-0.482	0.468	0.950	0.026	0.017
2	0	SCEE	-0.066	0.154	-0.516	0.451	0.967	-0.137	0.018
2	0	SFRE	-0.046	0.152	-0.465	0.545	1.010	0.265	0.018
2	0	TLIN	-0.036	0.149	-0.444	0.541	0.985	0.330	0.017
2	0	LLIN	-0.066	0.158	-0.572	0.385	0.957	-0.220	0.018
2	0	CARC	0.232	0.473	-0.477	1.328	1.805	0.583	0.055
2	0	IDEN	1.304	2.031	-1.487	5.095	6.582	0.361	0.234
2	0.05	CLIN	-0.141	0.172	-0.673	0.181	0.854	-0.932	0.020
2	0.05	SCEE	-0.160	0.178	-0.645	0.151	0.796	-0.875	0.021
2	0.05	SFRE	-0.175	0.162	-0.628	0.113	0.741	-0.924	0.019
2	0.05	TLIN	-0.162	0.159	-0.624	0.125	0.749	-0.916	0.018
2	0.05	LLIN	-0.117	0.190	-0.787	0.288	1.076	-0.927	0.022
2	0.05	CARC	0.146	0.486	-0.904	1.209	2.112	0.294	0.056
2	0.05	IDEN	1.261	2.032	-1.561	5.105	6.666	0.371	0.235
2	0.1	CLIN	-0.119	0.161	-0.414	0.469	0.883	0.726	0.019
2	0.1	SCEE	-0.140	0.162	-0.454	0.423	0.877	0.539	0.019
2	0.1	SFRE	-0.181	0.142	-0.435	0.297	0.732	0.559	0.016
2	0.1	TLIN	-0.169	0.141	-0.412	0.314	0.726	0.665	0.016
2	0.1	LLIN	-0.062	0.187	-0.422	0.647	1.069	0.780	0.022
2	0.1	CARC	0.164	0.471	-0.625	1.305	1.930	0.619	0.054
2	0.1	IDEN	1.277	2.021	-1.575	5.091	6.667	0.374	0.233
2	0.25	CLIN	-0.317	0.230	-0.822	0.183	1.004	-0.087	0.027
2	0.25	SCEE	-0.339	0.237	-0.866	0.189	1.055	-0.115	0.027
2	0.25	SFRE	-0.460	0.208	-0.907	0.045	0.952	0.021	0.024
2	0.25	TLIN	-0.454	0.205	-0.893	0.013	0.906	-0.031	0.024
2	0.25	LLIN	-0.160	0.260	-0.741	0.416	1.157	-0.152	0.030
2	0.25	CARC	-0.011	0.506	-0.757	1.106	1.863	0.568	0.058
2	0.25	IDEN	1.182	1.978	-1.417	4.942	6.358	0.391	0.228

Table A16: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for Various Levels of STD for Study 3 (60 Items, Average Item Discrimination 0.6)

Study	STD	Method	Mean	SD	Min	Max	Range	Skew	SE
3	0	CLIN	-0.018	0.140	-0.545	0.335	0.879	-0.679	0.016
3	0	SCEE	-0.025	0.138	-0.560	0.315	0.875	-0.827	0.016
3	0	SFRE	-0.023	0.125	-0.514	0.275	0.789	-0.969	0.014
3	0	TLIN	-0.015	0.126	-0.500	0.282	0.782	-0.892	0.015
3	0	LLIN	-0.022	0.160	-0.595	0.411	1.006	-0.449	0.018
3	0	CARC	0.645	0.344	0.011	1.535	1.524	0.670	0.040
3	0	IDEN	0.977	1.676	-1.540	4.415	5.955	0.422	0.194
3	0.05	CLIN	-0.052	0.113	-0.539	0.219	0.758	-0.855	0.013
3	0.05	SCEE	-0.060	0.116	-0.563	0.198	0.761	-0.936	0.013
3	0.05	SFRE	-0.076	0.110	-0.533	0.176	0.710	-0.776	0.013
3	0.05	TLIN	-0.069	0.109	-0.510	0.186	0.696	-0.696	0.013
3	0.05	LLIN	-0.033	0.122	-0.572	0.276	0.847	-0.872	0.014
3	0.05	CARC	0.618	0.312	0.124	1.451	1.328	0.873	0.036
3	0.05	IDEN	0.946	1.647	-1.358	4.295	5.652	0.483	0.190
3	0.1	CLIN	-0.101	0.158	-0.504	0.171	0.675	-0.555	0.018
3	0.1	SCEE	-0.109	0.162	-0.526	0.173	0.698	-0.574	0.019
3	0.1	SFRE	-0.148	0.146	-0.526	0.123	0.649	-0.592	0.017
3	0.1	TLIN	-0.141	0.148	-0.518	0.125	0.643	-0.568	0.017
3	0.1	LLIN	-0.055	0.173	-0.488	0.224	0.712	-0.516	0.020
3	0.1	CARC	0.572	0.370	0.039	1.541	1.502	0.910	0.043
3	0.1	IDEN	0.930	1.673	-1.421	4.309	5.730	0.499	0.193
3	0.25	CLIN	-0.206	0.279	-0.838	0.283	1.121	-0.301	0.032
3	0.25	SCEE	-0.215	0.283	-0.855	0.285	1.141	-0.316	0.033
3	0.25	SFRE	-0.327	0.241	-0.886	0.087	0.973	-0.410	0.028
3	0.25	TLIN	-0.321	0.245	-0.883	0.104	0.987	-0.394	0.028
3	0.25	LLIN	-0.076	0.320	-0.787	0.591	1.378	-0.210	0.037
3	0.25	CARC	0.479	0.486	-0.258	1.708	1.966	0.705	0.056
3	0.25	IDEN	0.874	1.691	-1.297	4.314	5.611	0.536	0.195

Table A17: Descriptive Statistics for the av.BIAS Pooled Over All Conditions for Various Levels of STD for Study 4 (60 Items, Average Item Discrimination 1.0)

Study	STD	Method	Mean	SD	Min	Max	Range	Skew	SE
4	0	CLIN	-0.012	0.109	-0.378	0.292	0.669	-0.403	0.013
4	0	SCEE	-0.033	0.113	-0.406	0.238	0.644	-0.453	0.013
4	0	SFRE	-0.032	0.107	-0.397	0.257	0.654	-0.344	0.012
4	0	TLIN	-0.013	0.104	-0.380	0.297	0.677	-0.291	0.012
4	0	LLIN	-0.011	0.116	-0.375	0.286	0.661	-0.503	0.013
4	0	CARC	0.773	0.475	0.011	1.898	1.887	0.667	0.055
4	0	IDEN	1.289	2.040	-1.564	5.002	6.566	0.329	0.236
4	0.05	CLIN	-0.044	0.113	-0.384	0.243	0.627	-0.590	0.013
4	0.05	SCEE	-0.066	0.122	-0.434	0.246	0.679	-0.606	0.014
4	0.05	SFRE	-0.079	0.118	-0.431	0.262	0.693	-0.588	0.014
4	0.05	TLIN	-0.062	0.109	-0.395	0.254	0.650	-0.593	0.013
4	0.05	LLIN	-0.025	0.121	-0.372	0.254	0.626	-0.524	0.014
4	0.05	CARC	0.740	0.481	0.000	1.877	1.878	0.678	0.055
4	0.05	IDEN	1.267	2.035	-1.682	4.983	6.664	0.349	0.235
4	0.1	CLIN	-0.084	0.140	-0.402	0.344	0.747	0.190	0.016
4	0.1	SCEE	-0.106	0.152	-0.437	0.313	0.750	0.003	0.018
4	0.1	SFRE	-0.141	0.140	-0.490	0.229	0.718	-0.092	0.016
4	0.1	TLIN	-0.124	0.130	-0.463	0.251	0.714	0.050	0.015
4	0.1	LLIN	-0.042	0.154	-0.347	0.446	0.793	0.295	0.018
4	0.1	CARC	0.708	0.524	-0.072	1.808	1.880	0.709	0.061
4	0.1	IDEN	1.239	2.034	-1.569	4.918	6.487	0.338	0.235
4	0.25	CLIN	-0.187	0.248	-0.831	0.513	1.344	-0.373	0.029
4	0.25	SCEE	-0.204	0.265	-0.879	0.534	1.412	-0.377	0.031
4	0.25	SFRE	-0.294	0.244	-0.925	0.392	1.317	-0.408	0.028
4	0.25	TLIN	-0.286	0.233	-0.896	0.378	1.274	-0.396	0.027
4	0.25	LLIN	-0.080	0.266	-0.761	0.659	1.420	-0.345	0.031
4	0.25	CARC	0.611	0.556	-0.293	1.892	2.185	0.648	0.064
4	0.25	IDEN	1.167	2.004	-1.484	4.913	6.397	0.380	0.231

APPENDIX B: DESCRIPTIVE STATISTICS FOR AVERAGE RMSD

Table B1: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for the Various Values of Test Length for Studies 1-4

Study	Length	Method	Mean	SD	Min	Max	Range	Skew	SE
1	30	CLIN	2.672	0.492	1.982	3.698	1.716	0.545	0.028
1	30	SCEE	2.658	0.507	1.981	3.719	1.738	0.532	0.029
1	30	SFRE	2.665	0.572	1.939	3.812	1.873	0.509	0.033
1	30	TLIN	2.683	0.548	1.952	3.770	1.818	0.529	0.032
1	30	LLIN	2.688	0.424	2.039	3.632	1.593	0.549	0.024
1	30	CARC	2.736	0.329	2.256	3.468	1.212	0.476	0.019
1	30	IDEN	3.181	0.669	2.252	4.912	2.660	0.762	0.039
2	30	CLIN	2.476	0.422	1.842	3.301	1.459	0.337	0.024
2	30	SCEE	2.402	0.460	1.748	3.305	1.557	0.531	0.027
2	30	SFRE	2.402	0.482	1.708	3.331	1.624	0.530	0.028
2	30	TLIN	2.480	0.441	1.817	3.321	1.505	0.345	0.025
2	30	LLIN	2.489	0.405	1.853	3.314	1.461	0.336	0.023
2	30	CARC	2.538	0.327	2.007	3.137	1.130	0.233	0.019
2	30	IDEN	3.264	1.011	2.006	5.707	3.701	0.798	0.058
3	60	CLIN	2.025	0.374	1.563	2.784	1.220	0.618	0.022
3	60	SCEE	1.996	0.397	1.523	2.818	1.295	0.660	0.023
3	60	SFRE	1.971	0.429	1.441	2.839	1.398	0.623	0.025
3	60	TLIN	2.004	0.401	1.497	2.803	1.306	0.610	0.023
3	60	LLIN	2.074	0.358	1.632	2.800	1.168	0.578	0.021
3	60	CARC	2.201	0.346	1.590	2.800	1.209	0.064	0.020
3	60	IDEN	2.677	0.801	1.627	4.730	3.104	0.793	0.046
4	60	CLIN	1.904	0.373	1.251	2.552	1.301	0.096	0.022
4	60	SCEE	1.788	0.395	1.308	2.539	1.231	0.583	0.023
4	60	SFRE	1.770	0.390	1.292	2.501	1.209	0.562	0.022
4	60	TLIN	1.892	0.369	1.258	2.535	1.277	0.074	0.021
4	60	LLIN	1.931	0.384	1.270	2.634	1.364	0.138	0.022
4	60	CARC	2.087	0.387	1.506	2.834	1.328	-0.021	0.022
4	60	IDEN	2.848	1.156	1.411	5.389	3.978	0.710	0.067

Table B2: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for the Various Levels of SMD for Study 1 (30 Items, Average Item Discrimination 0.6)

Study	SMD	Method	Mean	SD	Min	Max	Range	Skew	SE
1	0	CLIN	2.647	0.526	1.982	3.643	1.661	0.513	0.068
1	0	SCEE	2.655	0.525	1.981	3.650	1.669	0.505	0.068
1	0	SFRE	2.667	0.589	1.939	3.789	1.850	0.484	0.076
1	0	TLIN	2.661	0.584	1.952	3.768	1.816	0.492	0.075
1	0	LLIN	2.657	0.454	2.069	3.512	1.443	0.554	0.059
1	0	CARC	2.674	0.375	2.256	3.468	1.212	0.750	0.048
1	0	IDEN	2.657	0.368	2.253	3.429	1.176	0.738	0.047
1	-0.1	CLIN	2.668	0.521	2.055	3.558	1.503	0.532	0.067
1	-0.1	SCEE	2.672	0.523	2.050	3.589	1.540	0.512	0.068
1	-0.1	SFRE	2.680	0.585	1.991	3.610	1.619	0.496	0.076
1	-0.1	TLIN	2.679	0.576	2.022	3.589	1.567	0.518	0.074
1	-0.1	LLIN	2.684	0.452	2.146	3.608	1.462	0.563	0.058
1	-0.1	CARC	2.676	0.374	2.267	3.371	1.104	0.662	0.048
1	-0.1	IDEN	2.685	0.346	2.252	3.390	1.139	0.662	0.045
1	-0.5	CLIN	2.648	0.471	2.038	3.481	1.443	0.568	0.061
1	-0.5	SCEE	2.621	0.499	2.046	3.477	1.431	0.534	0.064
1	-0.5	SFRE	2.627	0.564	1.999	3.579	1.580	0.516	0.073
1	-0.5	TLIN	2.661	0.528	2.062	3.577	1.515	0.560	0.068
1	-0.5	LLIN	2.663	0.408	2.039	3.374	1.335	0.524	0.053
1	-0.5	CARC	2.727	0.295	2.334	3.283	0.949	0.510	0.038
1	-0.5	IDEN	3.413	0.196	3.017	3.726	0.709	0.180	0.025
1	-0.75	CLIN	2.721	0.421	2.228	3.449	1.221	0.591	0.054
1	-0.75	SCEE	2.652	0.484	2.012	3.470	1.458	0.476	0.063
1	-0.75	SFRE	2.652	0.548	2.005	3.507	1.502	0.465	0.071
1	-0.75	TLIN	2.726	0.478	2.203	3.492	1.289	0.573	0.062
1	-0.75	LLIN	2.743	0.357	2.325	3.493	1.168	0.590	0.046
1	-0.75	CARC	2.846	0.228	2.498	3.297	0.798	0.269	0.029
1	-0.75	IDEN	4.194	0.477	3.415	4.912	1.496	0.102	0.062
1	0.25	CLIN	2.678	0.523	2.066	3.698	1.632	0.562	0.068
1	0.25	SCEE	2.689	0.519	2.069	3.719	1.650	0.556	0.067
1	0.25	SFRE	2.699	0.588	1.969	3.812	1.843	0.509	0.076
1	0.25	TLIN	2.689	0.583	1.986	3.770	1.784	0.520	0.075
1	0.25	LLIN	2.692	0.449	2.136	3.632	1.496	0.581	0.058
1	0.25	CARC	2.758	0.331	2.271	3.352	1.081	0.624	0.043
1	0.25	IDEN	2.955	0.257	2.464	3.449	0.985	0.508	0.033

Table B3: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for the Various Levels of SMD for Study 2 (30 Items, Average Item Discrimination 1.0)

Study	SMD	Method	Mean	SD	Min	Max	Range	Skew	SE
2	0	CLIN	2.381	0.461	1.842	3.262	1.420	0.555	0.059
2	0	SCEE	2.386	0.465	1.850	3.296	1.446	0.544	0.060
2	0	SFRE	2.386	0.489	1.839	3.275	1.436	0.528	0.063
2	0	TLIN	2.385	0.481	1.817	3.256	1.440	0.539	0.062
2	0	LLIN	2.397	0.442	1.853	3.314	1.461	0.568	0.057
2	0	CARC	2.431	0.366	2.056	3.137	1.081	0.707	0.047
2	0	IDEN	2.413	0.354	2.006	3.084	1.079	0.714	0.046
2	-0.1	CLIN	2.394	0.439	1.907	3.244	1.337	0.561	0.057
2	-0.1	SCEE	2.383	0.458	1.865	3.232	1.368	0.514	0.059
2	-0.1	SFRE	2.379	0.482	1.856	3.234	1.378	0.514	0.062
2	-0.1	TLIN	2.396	0.459	1.901	3.240	1.339	0.558	0.059
2	-0.1	LLIN	2.412	0.423	1.916	3.295	1.379	0.562	0.055
2	-0.1	CARC	2.411	0.360	2.027	3.102	1.076	0.667	0.046
2	-0.1	IDEN	2.492	0.303	2.164	3.016	0.851	0.670	0.039
2	-0.5	CLIN	2.495	0.392	2.050	3.301	1.251	0.632	0.051
2	-0.5	SCEE	2.365	0.484	1.748	3.305	1.557	0.544	0.063
2	-0.5	SFRE	2.367	0.504	1.708	3.313	1.605	0.536	0.065
2	-0.5	TLIN	2.499	0.411	1.980	3.321	1.341	0.614	0.053
2	-0.5	LLIN	2.507	0.377	2.094	3.298	1.204	0.655	0.049
2	-0.5	CARC	2.570	0.269	2.221	3.135	0.915	0.602	0.035
2	-0.5	IDEN	3.742	0.218	3.299	4.244	0.946	-0.052	0.028
2	-0.75	CLIN	2.686	0.275	2.378	3.141	0.763	0.592	0.035
2	-0.75	SCEE	2.435	0.452	1.875	3.158	1.283	0.514	0.058
2	-0.75	SFRE	2.436	0.473	1.879	3.184	1.305	0.519	0.061
2	-0.75	TLIN	2.695	0.293	2.372	3.164	0.792	0.615	0.038
2	-0.75	LLIN	2.692	0.259	2.364	3.142	0.778	0.545	0.033
2	-0.75	CARC	2.733	0.185	2.405	3.132	0.727	0.136	0.024
2	-0.75	IDEN	4.904	0.527	3.988	5.707	1.719	-0.305	0.068
2	0.25	CLIN	2.422	0.453	1.885	3.213	1.328	0.533	0.058
2	0.25	SCEE	2.440	0.449	1.909	3.247	1.339	0.540	0.058
2	0.25	SFRE	2.443	0.475	1.888	3.331	1.444	0.549	0.061
2	0.25	TLIN	2.428	0.474	1.879	3.261	1.382	0.542	0.061
2	0.25	LLIN	2.437	0.434	1.913	3.288	1.375	0.533	0.056
2	0.25	CARC	2.545	0.322	2.007	3.134	1.127	0.602	0.042
2	0.25	IDEN	2.769	0.272	2.247	3.260	1.013	0.542	0.035

Table B4: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for the Various Levels of SMD for Study 3 (60 Items, Average Item Discrimination 0.6)

Study	SMD	Method	Mean	SD	Min	Max	Range	Skew	SE
3	0	CLIN	1.981	0.401	1.563	2.673	1.110	0.672	0.052
3	0	SCEE	1.986	0.406	1.574	2.701	1.127	0.662	0.052
3	0	SFRE	1.965	0.438	1.499	2.734	1.234	0.629	0.057
3	0	TLIN	1.963	0.431	1.497	2.710	1.213	0.646	0.056
3	0	LLIN	2.027	0.380	1.656	2.713	1.057	0.665	0.049
3	0	CARC	2.075	0.356	1.652	2.692	1.040	0.568	0.046
3	0	IDEN	2.024	0.377	1.627	2.650	1.023	0.601	0.049
3	-0.1	CLIN	1.995	0.403	1.576	2.784	1.208	0.698	0.052
3	-0.1	SCEE	1.996	0.410	1.570	2.818	1.248	0.693	0.053
3	-0.1	SFRE	1.971	0.444	1.509	2.839	1.330	0.644	0.057
3	-0.1	TLIN	1.973	0.434	1.501	2.803	1.302	0.664	0.056
3	-0.1	LLIN	2.044	0.381	1.632	2.800	1.168	0.680	0.049
3	-0.1	CARC	2.115	0.348	1.681	2.750	1.069	0.531	0.045
3	-0.1	IDEN	2.080	0.351	1.644	2.703	1.059	0.527	0.045
3	-0.5	CLIN	2.034	0.355	1.662	2.673	1.011	0.639	0.046
3	-0.5	SCEE	1.981	0.398	1.523	2.686	1.164	0.620	0.051
3	-0.5	SFRE	1.954	0.431	1.441	2.691	1.250	0.584	0.056
3	-0.5	TLIN	2.012	0.382	1.590	2.674	1.084	0.642	0.049
3	-0.5	LLIN	2.084	0.341	1.709	2.739	1.029	0.564	0.044
3	-0.5	CARC	2.280	0.262	1.806	2.669	0.863	-0.265	0.034
3	-0.5	IDEN	3.037	0.271	2.587	3.709	1.122	0.645	0.035
3	-0.75	CLIN	2.100	0.320	1.732	2.683	0.951	0.615	0.041
3	-0.75	SCEE	1.997	0.394	1.573	2.685	1.111	0.618	0.051
3	-0.75	SFRE	1.972	0.423	1.507	2.667	1.159	0.587	0.055
3	-0.75	TLIN	2.079	0.342	1.711	2.662	0.951	0.666	0.044
3	-0.75	LLIN	2.151	0.313	1.753	2.778	1.025	0.482	0.040
3	-0.75	CARC	2.444	0.248	2.035	2.800	0.765	-0.483	0.032
3	-0.75	IDEN	3.889	0.565	3.000	4.730	1.730	0.040	0.073
3	0.25	CLIN	2.015	0.385	1.623	2.685	1.062	0.652	0.050
3	0.25	SCEE	2.019	0.390	1.624	2.708	1.084	0.651	0.050
3	0.25	SFRE	1.995	0.421	1.553	2.704	1.151	0.613	0.054
3	0.25	TLIN	1.994	0.413	1.568	2.686	1.118	0.628	0.053
3	0.25	LLIN	2.064	0.369	1.634	2.735	1.101	0.641	0.048
3	0.25	CARC	2.091	0.358	1.590	2.780	1.189	0.566	0.046
3	0.25	IDEN	2.358	0.265	1.953	2.862	0.909	0.327	0.034

Table B5: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for the Various Levels of SMD for Study 4 (60 Items, Average Item Discrimination 1.0)

Study	SMD	Method	Mean	SD	Min	Max	Range	Skew	SE
4	0	CLIN	1.772	0.413	1.251	2.530	1.279	0.606	0.053
4	0	SCEE	1.775	0.411	1.308	2.515	1.207	0.594	0.053
4	0	SFRE	1.760	0.403	1.322	2.466	1.143	0.572	0.052
4	0	TLIN	1.760	0.407	1.258	2.483	1.225	0.593	0.053
4	0	LLIN	1.800	0.429	1.270	2.634	1.364	0.631	0.055
4	0	CARC	1.890	0.380	1.510	2.593	1.083	0.688	0.049
4	0	IDEN	1.815	0.405	1.411	2.525	1.114	0.683	0.052
4	-0.1	CLIN	1.783	0.405	1.361	2.552	1.191	0.613	0.052
4	-0.1	SCEE	1.766	0.416	1.322	2.539	1.217	0.600	0.054
4	-0.1	SFRE	1.748	0.413	1.292	2.501	1.209	0.574	0.053
4	-0.1	TLIN	1.770	0.403	1.330	2.535	1.204	0.603	0.052
4	-0.1	LLIN	1.811	0.416	1.389	2.607	1.217	0.626	0.054
4	-0.1	CARC	1.934	0.349	1.506	2.501	0.995	0.664	0.045
4	-0.1	IDEN	1.941	0.337	1.562	2.506	0.944	0.678	0.044
4	-0.5	CLIN	1.956	0.281	1.693	2.441	0.748	0.670	0.036
4	-0.5	SCEE	1.758	0.404	1.344	2.418	1.074	0.616	0.052
4	-0.5	SFRE	1.740	0.397	1.329	2.376	1.047	0.605	0.051
4	-0.5	TLIN	1.944	0.275	1.681	2.412	0.731	0.675	0.035
4	-0.5	LLIN	1.981	0.295	1.700	2.494	0.794	0.665	0.038
4	-0.5	CARC	2.204	0.209	1.775	2.637	0.862	-0.301	0.027
4	-0.5	IDEN	3.462	0.274	2.794	3.865	1.071	-0.091	0.035
4	-0.75	CLIN	2.190	0.163	1.908	2.517	0.609	0.357	0.021
4	-0.75	SCEE	1.822	0.381	1.399	2.446	1.047	0.542	0.049
4	-0.75	SFRE	1.802	0.375	1.392	2.414	1.023	0.524	0.048
4	-0.75	TLIN	2.179	0.157	1.907	2.483	0.577	0.422	0.020
4	-0.75	LLIN	2.214	0.178	1.925	2.610	0.685	0.322	0.023
4	-0.75	CARC	2.503	0.171	2.166	2.834	0.667	0.160	0.022
4	-0.75	IDEN	4.699	0.548	3.709	5.389	1.680	-0.363	0.071
4	0.25	CLIN	1.820	0.371	1.401	2.442	1.041	0.608	0.048
4	0.25	SCEE	1.820	0.372	1.387	2.442	1.056	0.589	0.048
4	0.25	SFRE	1.802	0.367	1.352	2.410	1.059	0.559	0.047
4	0.25	TLIN	1.805	0.368	1.374	2.417	1.043	0.585	0.047
4	0.25	LLIN	1.850	0.383	1.438	2.489	1.051	0.626	0.049
4	0.25	CARC	1.906	0.364	1.517	2.489	0.973	0.692	0.047
4	0.25	IDEN	2.322	0.227	1.930	2.734	0.804	0.539	0.029

Table B6: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for Various Levels of STD for Study 1 (30 Items, Average Item Discrimination 0.6)

Study	Discrimination	Method	Mean	SD	Min	Max	Range	Skew	SE
1	0.3	CLIN	3.335	0.113	3.063	3.698	0.635	0.501	0.011
1	0.3	SCEE	3.341	0.112	3.070	3.719	0.649	0.628	0.011
1	0.3	SFRE	3.434	0.113	3.169	3.812	0.643	0.483	0.011
1	0.3	TLIN	3.424	0.113	3.138	3.770	0.632	0.303	0.011
1	0.3	LLIN	3.251	0.124	2.961	3.632	0.671	0.813	0.012
1	0.3	CARC	3.163	0.106	2.820	3.468	0.648	-0.068	0.011
1	0.3	IDEN	3.313	0.211	2.922	3.899	0.977	0.763	0.021
1	0.6	CLIN	2.459	0.101	2.275	2.709	0.433	0.633	0.010
1	0.6	SCEE	2.453	0.105	2.295	2.705	0.410	0.875	0.011
1	0.6	SFRE	2.453	0.094	2.302	2.725	0.423	0.802	0.009
1	0.6	TLIN	2.465	0.088	2.302	2.690	0.388	0.637	0.009
1	0.6	LLIN	2.474	0.131	2.276	2.797	0.520	0.868	0.013
1	0.6	CARC	2.485	0.099	2.256	2.696	0.440	0.006	0.010
1	0.6	IDEN	3.002	0.659	2.252	4.351	2.099	0.778	0.066
1	0.9	CLIN	2.223	0.114	1.982	2.567	0.585	0.487	0.011
1	0.9	SCEE	2.180	0.100	1.981	2.569	0.589	1.013	0.010
1	0.9	SFRE	2.108	0.091	1.939	2.495	0.557	1.063	0.009
1	0.9	TLIN	2.162	0.102	1.952	2.487	0.535	0.470	0.010
1	0.9	LLIN	2.338	0.141	2.039	2.740	0.701	0.575	0.014
1	0.9	CARC	2.561	0.167	2.266	2.969	0.703	0.483	0.017
1	0.9	IDEN	3.228	0.906	2.269	4.912	2.642	0.749	0.091

Table B7: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for Various Levels of STD for Study 2 (30 Items, Average Item Discrimination 1.0)

Study	Discrimination	Method	Mean	SD	Min	Max	Range	Skew	SE
2	0.5	CLIN	3.023	0.085	2.857	3.301	0.444	0.635	0.008
2	0.5	SCEE	3.024	0.086	2.861	3.305	0.443	0.913	0.009
2	0.5	SFRE	3.055	0.088	2.890	3.331	0.442	0.747	0.009
2	0.5	TLIN	3.055	0.086	2.854	3.321	0.467	0.383	0.009
2	0.5	LLIN	3.010	0.094	2.837	3.314	0.476	1.112	0.009
2	0.5	CARC	2.940	0.083	2.780	3.137	0.357	0.413	0.008
2	0.5	IDEN	3.345	0.521	2.725	4.559	1.835	0.849	0.052
2	1	CLIN	2.286	0.140	2.081	2.610	0.529	0.716	0.014
2	1	SCEE	2.212	0.086	2.052	2.521	0.469	0.983	0.009
2	1	SFRE	2.209	0.080	2.072	2.537	0.465	1.014	0.008
2	1	TLIN	2.288	0.138	2.109	2.600	0.491	0.803	0.014
2	1	LLIN	2.299	0.147	2.066	2.637	0.572	0.573	0.015
2	1	CARC	2.319	0.153	2.007	2.815	0.808	0.561	0.015
2	1	IDEN	3.144	1.086	2.006	5.277	3.271	0.763	0.109
2	1.5	CLIN	2.118	0.210	1.842	2.665	0.823	0.925	0.021
2	1.5	SCEE	1.969	0.086	1.748	2.231	0.483	0.565	0.009
2	1.5	SFRE	1.943	0.079	1.708	2.154	0.446	0.431	0.008
2	1.5	TLIN	2.099	0.212	1.817	2.667	0.850	0.967	0.021
2	1.5	LLIN	2.159	0.210	1.853	2.721	0.868	0.837	0.021
2	1.5	CARC	2.354	0.216	2.027	2.804	0.777	0.602	0.022
2	1.5	IDEN	3.303	1.269	2.111	5.707	3.596	0.754	0.127

Table B8: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for Various Levels of STD for Study 3 (60 Items, Average Item Discrimination 0.6)

Study	Discrimination	Method	Mean	SD	Min	Max	Range	Skew	SE
3	0.3	CLIN	2.537	0.078	2.367	2.784	0.417	0.350	0.008
3	0.3	SCEE	2.545	0.080	2.380	2.818	0.438	0.591	0.008
3	0.3	SFRE	2.560	0.086	2.372	2.839	0.467	0.383	0.009
3	0.3	TLIN	2.553	0.084	2.355	2.803	0.448	0.127	0.008
3	0.3	LLIN	2.559	0.080	2.414	2.800	0.386	0.782	0.008
3	0.3	CARC	2.570	0.076	2.405	2.800	0.395	0.504	0.008
3	0.3	IDEN	2.769	0.272	2.399	3.474	1.075	0.984	0.027
3	0.6	CLIN	1.776	0.080	1.627	2.106	0.479	1.191	0.008
3	0.6	SCEE	1.758	0.075	1.615	2.085	0.469	1.288	0.008
3	0.6	SFRE	1.755	0.073	1.610	2.030	0.420	1.107	0.007
3	0.6	TLIN	1.774	0.078	1.636	2.064	0.428	1.155	0.008
3	0.6	LLIN	1.790	0.086	1.632	2.181	0.549	1.236	0.009
3	0.6	CARC	1.871	0.153	1.590	2.236	0.646	0.636	0.015
3	0.6	IDEN	2.450	0.827	1.627	3.953	2.327	0.705	0.083
3	0.9	CLIN	1.762	0.112	1.563	2.102	0.538	0.620	0.011
3	0.9	SCEE	1.685	0.075	1.523	1.888	0.366	0.782	0.007
3	0.9	SFRE	1.600	0.073	1.441	1.844	0.403	1.180	0.007
3	0.9	TLIN	1.685	0.111	1.497	2.005	0.508	0.646	0.011
3	0.9	LLIN	1.873	0.119	1.656	2.230	0.574	0.635	0.012
3	0.9	CARC	2.162	0.291	1.804	2.751	0.947	0.715	0.029
3	0.9	IDEN	2.814	1.049	1.707	4.730	3.023	0.700	0.105

Table B9: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for Various Levels of STD for Study 4 (60 Items, Average Item Discrimination 1.0)

Study	Discrimination	Method	Mean	SD	Min	Max	Range	Skew	SE
4	0.5	CLIN	2.349	0.064	2.191	2.552	0.361	0.492	0.006
4	0.5	SCEE	2.330	0.062	2.193	2.539	0.346	0.692	0.006
4	0.5	SFRE	2.303	0.062	2.118	2.501	0.383	0.239	0.006
4	0.5	TLIN	2.330	0.066	2.137	2.535	0.398	0.227	0.007
4	0.5	LLIN	2.395	0.068	2.230	2.634	0.403	0.882	0.007
4	0.5	CARC	2.421	0.075	2.211	2.637	0.426	-0.235	0.008
4	0.5	IDEN	2.914	0.619	2.213	4.169	1.957	0.842	0.062
4	1	CLIN	1.707	0.184	1.458	2.202	0.745	0.944	0.018
4	1	SCEE	1.593	0.074	1.458	1.830	0.372	0.766	0.007
4	1	SFRE	1.591	0.070	1.471	1.787	0.316	0.646	0.007
4	1	TLIN	1.708	0.182	1.476	2.192	0.716	0.969	0.018
4	1	LLIN	1.714	0.185	1.460	2.222	0.761	0.914	0.018
4	1	CARC	1.848	0.289	1.506	2.560	1.054	0.943	0.029
4	1	IDEN	2.719	1.254	1.411	4.938	3.528	0.682	0.125
4	1.5	CLIN	1.656	0.283	1.251	2.219	0.968	0.747	0.028
4	1.5	SCEE	1.441	0.076	1.308	1.716	0.408	0.941	0.008
4	1.5	SFRE	1.417	0.073	1.292	1.685	0.393	1.056	0.007
4	1.5	TLIN	1.637	0.283	1.258	2.209	0.951	0.751	0.028
4	1.5	LLIN	1.685	0.282	1.270	2.253	0.983	0.736	0.028
4	1.5	CARC	1.992	0.428	1.510	2.834	1.324	0.719	0.043
4	1.5	IDEN	2.910	1.434	1.413	5.389	3.976	0.642	0.143

Table B10: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for Various Sample Sizes for Study 1 (30 Items, Average Item Discrimination 0.6)

Study	Size	Method	Mean	SD	Min	Max	Range	Skew	SE
1	25	CLIN	2.726	0.482	2.038	3.698	1.660	0.480	0.062
1	25	SCEE	2.726	0.495	2.012	3.719	1.707	0.428	0.064
1	25	SFRE	2.719	0.559	1.969	3.812	1.843	0.425	0.072
1	25	TLIN	2.720	0.539	1.986	3.770	1.784	0.473	0.070
1	25	LLIN	2.792	0.425	2.039	3.632	1.593	0.437	0.055
1	25	CARC	2.758	0.328	2.282	3.428	1.146	0.539	0.042
1	25	IDEN	3.172	0.680	2.328	4.912	2.584	0.870	0.088
1	50	CLIN	2.703	0.510	1.998	3.643	1.644	0.560	0.066
1	50	SCEE	2.694	0.521	2.000	3.650	1.649	0.572	0.067
1	50	SFRE	2.692	0.588	1.948	3.789	1.841	0.550	0.076
1	50	TLIN	2.707	0.567	1.952	3.768	1.816	0.550	0.073
1	50	LLIN	2.732	0.438	2.086	3.510	1.423	0.530	0.057
1	50	CARC	2.756	0.349	2.267	3.468	1.201	0.478	0.045
1	50	IDEN	3.189	0.659	2.252	4.781	2.530	0.621	0.085
1	100	CLIN	2.645	0.488	2.029	3.520	1.491	0.543	0.063
1	100	SCEE	2.627	0.504	2.037	3.515	1.478	0.553	0.065
1	100	SFRE	2.640	0.569	1.949	3.627	1.677	0.517	0.073
1	100	TLIN	2.662	0.545	1.961	3.626	1.665	0.521	0.070
1	100	LLIN	2.645	0.413	2.136	3.388	1.251	0.565	0.053
1	100	CARC	2.719	0.324	2.256	3.271	1.015	0.426	0.042
1	100	IDEN	3.170	0.674	2.253	4.856	2.603	0.755	0.087
1	200	CLIN	2.648	0.494	1.982	3.481	1.499	0.533	0.064
1	200	SCEE	2.626	0.513	1.981	3.477	1.496	0.525	0.066
1	200	SFRE	2.640	0.579	1.939	3.579	1.640	0.495	0.075
1	200	TLIN	2.666	0.553	1.963	3.577	1.614	0.510	0.071
1	200	LLIN	2.641	0.416	2.069	3.366	1.297	0.572	0.054
1	200	CARC	2.726	0.326	2.266	3.283	1.017	0.439	0.042
1	200	IDEN	3.188	0.675	2.269	4.893	2.624	0.753	0.087
1	400	CLIN	2.641	0.493	2.055	3.453	1.398	0.564	0.064
1	400	SCEE	2.617	0.512	2.050	3.457	1.407	0.557	0.066
1	400	SFRE	2.633	0.576	1.991	3.559	1.568	0.519	0.074
1	400	TLIN	2.661	0.551	2.022	3.555	1.533	0.535	0.071
1	400	LLIN	2.629	0.417	2.146	3.328	1.182	0.609	0.054
1	400	CARC	2.722	0.326	2.331	3.253	0.923	0.422	0.042
1	400	IDEN	3.185	0.680	2.353	4.837	2.485	0.725	0.088

Table B11: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for Various Sample Sizes for Study 2 (30 Items, Average Item Discrimination 1.0)

Study	Size	Method	Mean	SD	Min	Max	Range	Skew	SE
2	25	CLIN	2.514	0.427	1.883	3.262	1.379	0.380	0.055
2	25	SCEE	2.463	0.462	1.748	3.296	1.548	0.492	0.060
2	25	SFRE	2.448	0.484	1.708	3.331	1.624	0.491	0.062
2	25	TLIN	2.505	0.447	1.817	3.261	1.444	0.352	0.058
2	25	LLIN	2.564	0.412	1.992	3.314	1.322	0.413	0.053
2	25	CARC	2.553	0.336	2.007	3.137	1.130	0.300	0.043
2	25	IDEN	3.255	1.028	2.006	5.707	3.701	0.832	0.133
2	50	CLIN	2.486	0.416	1.874	3.301	1.427	0.370	0.054
2	50	SCEE	2.416	0.458	1.875	3.305	1.430	0.560	0.059
2	50	SFRE	2.412	0.479	1.845	3.313	1.468	0.556	0.062
2	50	TLIN	2.488	0.435	1.861	3.321	1.460	0.374	0.056
2	50	LLIN	2.506	0.397	1.908	3.298	1.390	0.371	0.051
2	50	CARC	2.536	0.326	2.057	3.135	1.078	0.280	0.042
2	50	IDEN	3.253	1.007	2.078	5.597	3.519	0.770	0.130
2	100	CLIN	2.475	0.432	1.842	3.125	1.283	0.329	0.056
2	100	SCEE	2.401	0.468	1.825	3.156	1.331	0.510	0.060
2	100	SFRE	2.406	0.491	1.800	3.206	1.406	0.516	0.063
2	100	TLIN	2.484	0.452	1.853	3.181	1.329	0.350	0.058
2	100	LLIN	2.478	0.411	1.853	3.105	1.252	0.314	0.053
2	100	CARC	2.542	0.329	2.065	3.064	0.999	0.211	0.042
2	100	IDEN	3.268	1.010	2.083	5.493	3.410	0.760	0.130
2	200	CLIN	2.458	0.429	1.885	3.086	1.201	0.282	0.055
2	200	SCEE	2.373	0.463	1.859	3.092	1.233	0.527	0.060
2	200	SFRE	2.381	0.488	1.856	3.148	1.292	0.525	0.063
2	200	TLIN	2.468	0.449	1.873	3.139	1.266	0.305	0.058
2	200	LLIN	2.456	0.409	1.900	3.057	1.157	0.261	0.053
2	200	CARC	2.534	0.333	2.027	3.027	1.000	0.143	0.043
2	200	IDEN	3.275	1.025	2.126	5.531	3.405	0.759	0.132
2	400	CLIN	2.446	0.419	1.881	3.077	1.196	0.297	0.054
2	400	SCEE	2.355	0.456	1.850	3.061	1.210	0.532	0.059
2	400	SFRE	2.364	0.481	1.832	3.116	1.285	0.521	0.062
2	400	TLIN	2.457	0.438	1.869	3.134	1.264	0.310	0.057
2	400	LLIN	2.441	0.398	1.904	3.028	1.125	0.289	0.051
2	400	CARC	2.524	0.321	2.069	3.016	0.947	0.194	0.041
2	400	IDEN	3.268	1.017	2.143	5.539	3.396	0.784	0.131

Table B12: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for Various Sample Sizes for Study 3 (60 Items, Average Item Discrimination 0.6)

Study	Size	Method	Mean	SD	Min	Max	Range	Skew	SE
3	25	CLIN	2.075	0.376	1.576	2.784	1.208	0.589	0.049
3	25	SCEE	2.060	0.393	1.573	2.818	1.245	0.661	0.051
3	25	SFRE	2.030	0.422	1.509	2.839	1.330	0.630	0.054
3	25	TLIN	2.046	0.403	1.501	2.803	1.302	0.581	0.052
3	25	LLIN	2.145	0.363	1.683	2.800	1.117	0.548	0.047
3	25	CARC	2.216	0.346	1.652	2.800	1.147	0.187	0.045
3	25	IDEN	2.687	0.806	1.627	4.631	3.005	0.723	0.104
3	50	CLIN	2.034	0.380	1.563	2.667	1.103	0.601	0.049
3	50	SCEE	2.008	0.403	1.574	2.691	1.117	0.646	0.052
3	50	SFRE	1.983	0.435	1.512	2.714	1.202	0.616	0.056
3	50	TLIN	2.012	0.408	1.497	2.709	1.212	0.602	0.053
3	50	LLIN	2.085	0.365	1.634	2.712	1.078	0.557	0.047
3	50	CARC	2.201	0.359	1.590	2.725	1.134	0.034	0.046
3	50	IDEN	2.665	0.813	1.631	4.730	3.099	0.828	0.105
3	100	CLIN	2.016	0.378	1.617	2.636	1.019	0.630	0.049
3	100	SCEE	1.984	0.403	1.523	2.646	1.124	0.648	0.052
3	100	SFRE	1.959	0.436	1.441	2.678	1.237	0.607	0.056
3	100	TLIN	1.995	0.407	1.549	2.674	1.125	0.614	0.053
3	100	LLIN	2.062	0.361	1.695	2.641	0.946	0.603	0.047
3	100	CARC	2.201	0.349	1.729	2.739	1.010	0.077	0.045
3	100	IDEN	2.678	0.805	1.646	4.623	2.978	0.776	0.104
3	200	CLIN	2.005	0.371	1.591	2.592	1.002	0.594	0.048
3	200	SCEE	1.969	0.395	1.570	2.600	1.031	0.660	0.051
3	200	SFRE	1.949	0.429	1.499	2.623	1.124	0.620	0.055
3	200	TLIN	1.989	0.400	1.522	2.619	1.096	0.590	0.052
3	200	LLIN	2.044	0.352	1.632	2.605	0.973	0.549	0.045
3	200	CARC	2.194	0.343	1.681	2.659	0.978	-0.012	0.044
3	200	IDEN	2.680	0.802	1.644	4.619	2.975	0.765	0.104
3	400	CLIN	1.995	0.371	1.620	2.557	0.937	0.626	0.048
3	400	SCEE	1.958	0.396	1.573	2.565	0.992	0.665	0.051
3	400	SFRE	1.937	0.430	1.507	2.600	1.092	0.618	0.056
3	400	TLIN	1.979	0.400	1.558	2.597	1.039	0.611	0.052
3	400	LLIN	2.033	0.351	1.656	2.553	0.897	0.585	0.045
3	400	CARC	2.193	0.345	1.682	2.686	1.004	0.027	0.045
3	400	IDEN	2.677	0.806	1.658	4.619	2.960	0.794	0.104

Table B13: Descriptive Statistics for the av.RMSD Pooled Over All Conditions for Various Sample Sizes for Study 4 (60 Items, Average Item Discrimination 1.0)

Study	Size	Method	Mean	SD	Min	Max	Range	Skew	SE
4	25	CLIN	1.926	0.369	1.251	2.552	1.301	0.146	0.048
4	25	SCEE	1.835	0.389	1.308	2.539	1.231	0.555	0.050
4	25	SFRE	1.809	0.382	1.292	2.501	1.209	0.535	0.049
4	25	TLIN	1.907	0.366	1.258	2.535	1.277	0.128	0.047
4	25	LLIN	1.969	0.378	1.270	2.634	1.364	0.194	0.049
4	25	CARC	2.067	0.374	1.506	2.717	1.211	0.028	0.048
4	25	IDEN	2.815	1.155	1.411	5.353	3.943	0.725	0.149
4	50	CLIN	1.918	0.373	1.361	2.442	1.081	0.076	0.048
4	50	SCEE	1.811	0.397	1.341	2.442	1.101	0.558	0.051
4	50	SFRE	1.791	0.392	1.322	2.410	1.088	0.529	0.051
4	50	TLIN	1.905	0.372	1.341	2.421	1.080	0.044	0.048
4	50	LLIN	1.947	0.382	1.396	2.494	1.098	0.132	0.049
4	50	CARC	2.094	0.389	1.534	2.834	1.300	0.034	0.050
4	50	IDEN	2.846	1.160	1.481	5.389	3.908	0.715	0.150
4	100	CLIN	1.902	0.386	1.374	2.443	1.069	0.094	0.050
4	100	SCEE	1.780	0.406	1.338	2.419	1.081	0.606	0.052
4	100	SFRE	1.764	0.400	1.326	2.402	1.076	0.589	0.052
4	100	TLIN	1.892	0.380	1.363	2.438	1.075	0.077	0.049
4	100	LLIN	1.925	0.398	1.383	2.462	1.079	0.127	0.051
4	100	CARC	2.090	0.397	1.517	2.765	1.248	-0.067	0.051
4	100	IDEN	2.860	1.172	1.474	5.361	3.887	0.667	0.151
4	200	CLIN	1.891	0.373	1.367	2.424	1.058	0.108	0.048
4	200	SCEE	1.762	0.398	1.322	2.346	1.024	0.587	0.051
4	200	SFRE	1.749	0.393	1.299	2.336	1.037	0.565	0.051
4	200	TLIN	1.882	0.369	1.347	2.412	1.065	0.084	0.048
4	200	LLIN	1.912	0.383	1.392	2.460	1.069	0.150	0.049
4	200	CARC	2.089	0.389	1.560	2.761	1.201	-0.056	0.050
4	200	IDEN	2.863	1.165	1.499	5.310	3.811	0.687	0.150
4	400	CLIN	1.882	0.374	1.361	2.430	1.070	0.055	0.048
4	400	SCEE	1.752	0.394	1.332	2.375	1.043	0.589	0.051
4	400	SFRE	1.738	0.390	1.308	2.361	1.054	0.566	0.050
4	400	TLIN	1.873	0.369	1.348	2.423	1.075	0.035	0.048
4	400	LLIN	1.903	0.385	1.382	2.461	1.079	0.093	0.050
4	400	CARC	2.095	0.398	1.566	2.793	1.227	-0.045	0.051
4	400	IDEN	2.855	1.168	1.501	5.326	3.826	0.685	0.151

Table B14: Descriptive Statistics for the av.RMSD Over All Conditions for Various Levels of STD for Study 1 (30 Items, Average Item Discrimination 0.6)

Study	STD	Method	Mean	SD	Min	Max	Range	Skew	SE
1	0	CLIN	2.679	0.479	2.038	3.558	1.521	0.502	0.055
1	0	SCEE	2.662	0.497	2.046	3.589	1.543	0.484	0.057
1	0	SFRE	2.662	0.562	1.969	3.631	1.662	0.460	0.065
1	0	TLIN	2.680	0.536	1.986	3.614	1.628	0.491	0.062
1	0	LLIN	2.705	0.418	2.039	3.608	1.569	0.494	0.048
1	0	CARC	2.745	0.315	2.340	3.428	1.089	0.520	0.036
1	0	IDEN	3.209	0.677	2.333	4.904	2.571	0.739	0.078
1	0.05	CLIN	2.670	0.480	1.998	3.558	1.560	0.523	0.055
1	0.05	SCEE	2.651	0.494	2.000	3.508	1.508	0.511	0.057
1	0.05	SFRE	2.656	0.558	1.948	3.568	1.620	0.491	0.064
1	0.05	TLIN	2.675	0.536	1.952	3.574	1.622	0.507	0.062
1	0.05	LLIN	2.689	0.415	2.086	3.558	1.472	0.524	0.048
1	0.05	CARC	2.737	0.322	2.282	3.323	1.040	0.404	0.037
1	0.05	IDEN	3.196	0.677	2.311	4.912	2.600	0.766	0.078
1	0.1	CLIN	2.681	0.511	2.072	3.698	1.626	0.586	0.059
1	0.1	SCEE	2.666	0.529	2.012	3.719	1.707	0.561	0.061
1	0.1	SFRE	2.672	0.590	1.995	3.812	1.817	0.534	0.068
1	0.1	TLIN	2.691	0.566	2.013	3.770	1.757	0.556	0.065
1	0.1	LLIN	2.694	0.448	2.136	3.632	1.496	0.610	0.052
1	0.1	CARC	2.742	0.340	2.256	3.394	1.138	0.527	0.039
1	0.1	IDEN	3.184	0.665	2.253	4.837	2.584	0.682	0.077
1	0.25	CLIN	2.659	0.505	1.982	3.643	1.661	0.527	0.058
1	0.25	SCEE	2.653	0.519	1.981	3.650	1.669	0.524	0.060
1	0.25	SFRE	2.669	0.587	1.939	3.789	1.850	0.507	0.068
1	0.25	TLIN	2.688	0.565	1.961	3.768	1.807	0.514	0.065
1	0.25	LLIN	2.664	0.422	2.069	3.496	1.427	0.516	0.049
1	0.25	CARC	2.719	0.345	2.266	3.468	1.202	0.442	0.040
1	0.25	IDEN	3.135	0.668	2.252	4.870	2.618	0.816	0.077

Table B15: Descriptive Statistics for the av.RMSD Over All Conditions for Various Levels of STD for Study 2 (30 Items, Average Item Discrimination 1.0)

Study	STD	Method	Mean	SD	Min	Max	Range	Skew	SE
2	0	CLIN	2.497	0.409	1.963	3.262	1.299	0.345	0.047
2	0	SCEE	2.413	0.449	1.748	3.296	1.548	0.521	0.052
2	0	SFRE	2.405	0.475	1.708	3.331	1.624	0.510	0.055
2	0	TLIN	2.492	0.430	1.923	3.261	1.337	0.350	0.050
2	0	LLIN	2.520	0.389	1.991	3.314	1.323	0.341	0.045
2	0	CARC	2.559	0.321	2.134	3.137	1.003	0.219	0.037
2	0	IDEN	3.305	1.030	2.151	5.597	3.446	0.760	0.119
2	0.05	CLIN	2.470	0.422	1.886	3.244	1.358	0.420	0.049
2	0.05	SCEE	2.394	0.461	1.853	3.232	1.380	0.553	0.053
2	0.05	SFRE	2.390	0.482	1.832	3.234	1.403	0.554	0.056
2	0.05	TLIN	2.470	0.440	1.873	3.240	1.367	0.412	0.051
2	0.05	LLIN	2.484	0.406	1.907	3.295	1.388	0.438	0.047
2	0.05	CARC	2.542	0.326	2.108	3.102	0.995	0.274	0.038
2	0.05	IDEN	3.278	1.017	2.126	5.557	3.431	0.793	0.117
2	0.1	CLIN	2.463	0.423	1.874	3.301	1.427	0.268	0.049
2	0.1	SCEE	2.396	0.456	1.867	3.305	1.438	0.465	0.053
2	0.1	SFRE	2.395	0.476	1.839	3.313	1.473	0.467	0.055
2	0.1	TLIN	2.466	0.441	1.817	3.321	1.505	0.278	0.051
2	0.1	LLIN	2.477	0.407	1.904	3.298	1.395	0.265	0.047
2	0.1	CARC	2.531	0.325	2.007	3.135	1.129	0.135	0.037
2	0.1	IDEN	3.265	1.035	2.078	5.707	3.629	0.808	0.119
2	0.25	CLIN	2.473	0.444	1.842	3.213	1.371	0.313	0.051
2	0.25	SCEE	2.405	0.481	1.850	3.226	1.376	0.546	0.055
2	0.25	SFRE	2.418	0.505	1.844	3.261	1.417	0.541	0.058
2	0.25	TLIN	2.494	0.462	1.853	3.202	1.349	0.317	0.053
2	0.25	LLIN	2.476	0.424	1.853	3.288	1.435	0.314	0.049
2	0.25	CARC	2.519	0.342	2.056	3.088	1.032	0.300	0.039
2	0.25	IDEN	3.207	0.977	2.006	5.467	3.461	0.766	0.113

Table B16: Descriptive Statistics for the av.RMSD Over All Conditions for Various Levels of STD for Study 3 (60 Items, Average Item Discrimination 0.6)

Study	STD	Method	Mean	SD	Min	Max	Range	Skew	SE
3	0	CLIN	2.019	0.351	1.576	2.685	1.109	0.623	0.040
3	0	SCEE	1.989	0.375	1.573	2.708	1.134	0.645	0.043
3	0	SFRE	1.958	0.403	1.509	2.704	1.195	0.612	0.047
3	0	TLIN	1.990	0.377	1.501	2.686	1.186	0.610	0.043
3	0	LLIN	2.075	0.339	1.683	2.778	1.095	0.582	0.039
3	0	CARC	2.204	0.342	1.662	2.800	1.138	0.128	0.039
3	0	IDEN	2.692	0.809	1.638	4.730	3.092	0.814	0.093
3	0.05	CLIN	2.018	0.385	1.563	2.688	1.125	0.641	0.044
3	0.05	SCEE	1.986	0.409	1.574	2.710	1.136	0.675	0.047
3	0.05	SFRE	1.962	0.439	1.499	2.751	1.252	0.637	0.051
3	0.05	TLIN	1.995	0.412	1.497	2.725	1.228	0.625	0.048
3	0.05	LLIN	2.068	0.371	1.656	2.735	1.079	0.625	0.043
3	0.05	CARC	2.201	0.355	1.652	2.780	1.127	0.094	0.041
3	0.05	IDEN	2.674	0.799	1.627	4.612	2.985	0.726	0.092
3	0.1	CLIN	2.029	0.382	1.606	2.784	1.178	0.610	0.044
3	0.1	SCEE	2.000	0.405	1.523	2.818	1.295	0.656	0.047
3	0.1	SFRE	1.975	0.433	1.441	2.839	1.398	0.612	0.050
3	0.1	TLIN	2.007	0.406	1.543	2.803	1.260	0.601	0.047
3	0.1	LLIN	2.078	0.369	1.676	2.800	1.124	0.578	0.043
3	0.1	CARC	2.207	0.357	1.682	2.750	1.068	0.044	0.041
3	0.1	IDEN	2.681	0.817	1.631	4.631	3.000	0.763	0.094
3	0.25	CLIN	2.033	0.385	1.591	2.662	1.071	0.553	0.044
3	0.25	SCEE	2.009	0.407	1.570	2.701	1.131	0.615	0.047
3	0.25	SFRE	1.991	0.446	1.507	2.734	1.226	0.574	0.051
3	0.25	TLIN	2.023	0.417	1.550	2.709	1.159	0.551	0.048
3	0.25	LLIN	2.075	0.359	1.632	2.650	1.017	0.487	0.041
3	0.25	CARC	2.192	0.339	1.590	2.751	1.160	-0.023	0.039
3	0.25	IDEN	2.663	0.794	1.644	4.619	2.974	0.820	0.092

Table B17: Descriptive Statistics for the av.RMSD Over All Conditions for Various Levels of STD for Study 4 (60 Items, Average Item Discrimination 1.0)

Study	STD	Method	Mean	SD	Min	Max	Range	Skew	SE
4	0	CLIN	1.909	0.365	1.367	2.441	1.074	0.112	0.042
4	0	SCEE	1.788	0.392	1.322	2.418	1.096	0.571	0.045
4	0	SFRE	1.768	0.386	1.299	2.376	1.077	0.555	0.045
4	0	TLIN	1.893	0.361	1.347	2.421	1.073	0.095	0.042
4	0	LLIN	1.937	0.376	1.391	2.494	1.104	0.143	0.043
4	0	CARC	2.111	0.397	1.601	2.834	1.233	0.000	0.046
4	0	IDEN	2.872	1.181	1.503	5.389	3.886	0.686	0.136
4	0.05	CLIN	1.912	0.362	1.375	2.427	1.052	0.073	0.042
4	0.05	SCEE	1.790	0.390	1.365	2.421	1.055	0.546	0.045
4	0.05	SFRE	1.772	0.386	1.349	2.410	1.061	0.527	0.045
4	0.05	TLIN	1.897	0.361	1.370	2.404	1.034	0.055	0.042
4	0.05	LLIN	1.940	0.372	1.383	2.501	1.118	0.107	0.043
4	0.05	CARC	2.103	0.383	1.549	2.793	1.244	-0.040	0.044
4	0.05	IDEN	2.867	1.168	1.411	5.368	3.957	0.721	0.135
4	0.1	CLIN	1.908	0.372	1.361	2.530	1.169	0.111	0.043
4	0.1	SCEE	1.794	0.391	1.341	2.515	1.174	0.583	0.045
4	0.1	SFRE	1.773	0.383	1.322	2.443	1.121	0.555	0.044
4	0.1	TLIN	1.894	0.366	1.341	2.483	1.142	0.084	0.042
4	0.1	LLIN	1.936	0.387	1.382	2.634	1.251	0.160	0.045
4	0.1	CARC	2.093	0.384	1.586	2.732	1.146	-0.060	0.044
4	0.1	IDEN	2.857	1.158	1.486	5.297	3.811	0.672	0.134
4	0.25	CLIN	1.888	0.397	1.251	2.552	1.301	0.102	0.046
4	0.25	SCEE	1.780	0.415	1.308	2.539	1.231	0.600	0.048
4	0.25	SFRE	1.768	0.411	1.292	2.501	1.209	0.573	0.048
4	0.25	TLIN	1.883	0.395	1.258	2.535	1.277	0.072	0.046
4	0.25	LLIN	1.912	0.406	1.270	2.607	1.337	0.149	0.047
4	0.25	CARC	2.042	0.387	1.506	2.746	1.240	0.010	0.045
4	0.25	IDEN	2.795	1.140	1.413	5.299	3.886	0.715	0.132