CONSTRUCT CONTINUITY IN THE PRESENCE OF MULTIDIMENSIONALITY

by

Dorota Staniewska

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Education

written under the direction of

Douglas Penfield

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2009

ABSTRACT OF THE DISSERTATION

Construct Continuity in the Presence of Multidimensionality

By Dorota Staniewska

Dissertation Director:

Douglas Penfield

Unidimensionality – a condition, under which only one dominant construct is being measured by the test, is a fundamental assumption of most modern day psychometric models. However, some tests are multidimensional by design. A test, for instance, might measure physics, biology and chemistry subscales combined to measure a "general science" composite. The relative magnitudes of those subscales sometimes shift from administration to administration, which results in an altered composite. This study examined the conditions under which two different forms of a multidimensional test measure the same composite construct to a degree that allows them to be equated, i.e. used interchangeably.

IRT true-score equating was used in a simulation study to assess the closeness of the scores on the forms. Conditions examined included the correlations between subscales, varying number of items per subscale form to form, and different subpopulation ability estimates on the subscales. Differences in the equating errors due to generating model (1PL or 3PL) were also examined. A way of calculating a

unidimensional composite from a two-dimensional ability was devised and compared to the unidimensional composite obtained from Parscale.

It was found that in general, the errors increase with decreasing correlation between traits and increased divergence of the two forms to be equated, with the later being the main predictor of the equating errors. However, the magnitude of those errors was small for the population as a whole especially when all examinee abilities are drawn from the same distribution. It was concluded that IRT true score equating is relatively robust to multidimensionality for the conditions examined, especially if the overall population score is desired. However, when accurate estimate of the equated score for individuals at the extremes of the population is needed, or whenever population abilities are drawn from more than one distribution, the unidimensional true score equating functions well only for very similar forms and with high correlations between traits.

Table of Contents

List of Tables

List of figures

1 Introduction

1.1 Background

Unidimensionality – a condition under which only one dominant construct is being measured by the test, is a fundamental assumption of most modern day psychometric models. However, in practice, all cognitive constructs are multidimensional to some extent. For instance, a general science test can consist of chemistry, physics and biology subscales, each of which, while correlated with the others, might be considered a separate mini-test measuring a different skill. The way the subscales are combined to measure examinee or subgroup mastery of the material and, in subsequent test administrations, trend, is critical to the validity of the assessment.

Considerable research has gone into assessing if a test shows evidence of multidimensionality when it was not purposefully written to, with both parametric (e.g. factor analysis) and nonparametric (e.g. conditional covariance) methods applied to identify the clusters of items that measure the same ability and to examine the correlations among those clusters. This dissertation explores multidimensionality from a different viewpoint by considering a situation in which the multidimensional structure of the test is known, and the validity of the construct continuity assumption between subsequent assessment forms needs to be checked. Stated another way – how different can two multidimensional tests be and still measure the same overall ability.

Most assessments consist of separate subscales (dimensions) which are combined together to form a composite. For instance, the (new) SAT verbal has reading comprehension, sentence completion and paragraph-length critical reading items; the

GRE writing test consists of an issue task and an argument task; the NAEP Reading assessment measures reading for literary experience, reading for information and reading to perform a task. Those dimensions are combined to measure verbal, writing and reading composite for the three tests respectively. But the relative magnitudes of those subscales sometimes shift from administration to administration, which results in an altered composite. A passage measuring one of the subscales on a reading test could show Differential Item Functioning (DIF)[1] and might need to be removed, thus shifting the composite towards the remaining subscales. Or one of the subscales might be removed and replaced by a completely new one, as was the case with the (old) SAT verbal analogy items which were replaced by reading items on the (new) SAT verbal test. More commonly, a test is designed to certain specifications which cannot be fulfilled exactly year after year. The number of possible questions referring to a passage on a reading assessment for instance will depend on the passage. If a portion or all items are released, a different passage might consist of a different number of questions.

1.2 Statement of the problem

It is important to know when the construct measured by the test diverged from what it was originally designed to be. For one, the equated scores of examinees on a test are assumed to be equivalent even if they took the test at different administrations. A 2300 on the (new) SAT is intended to have the same meaning for an examinee that took the December form as it does for an examinee that took the April form. This assumption cannot be made if the forms are measuring different things! Secondly, any increase in

[1] That is – conditioned on ability (usually approximated by a function of raw score and model-specific parameters) at least one group performs worse than another group on an item.

examinee or group scores from administration to administration needs to be due to an increase in proficiency, not to change in the construct measured. If the conditions of measurement (format, timing, availability of accommodations for disabled students, etc) have changed between administrations, additional psychometric adjustments may be required and are usually employed. However, potential dimensional inconsistencies between administrations are rarely considered in this adjusting even though the validity of the assessment is at stake.

A slight shift in the composite between administrations is probably unavoidable with items being released (or replaced with other items). However, it is desirable to identify when the construct measured by the test has shifted too much to reasonably assume that the original construct is still being measured. One could speculate that if, in the extreme case all items from one subscale were removed, the "same construct" argument obviously could not be made.

This dissertation examines under what circumstances equating is no longer possible because of the shift in the composite described above when multidimensionality is assumed. It's assumed throughout that once equating fails (i.e. the scores between different administrations for forms of the test are not comparable), the same construct has not been measured.

1.3 Research Questions/Hypotheses.

The following research questions were explored:

1. What influences the magnitude of the equating errors for multidimensional test forms:

   a. Correlation between dimensions – it was hypothesized that the higher correlation between the subscales, the more difficult it was going to be to distinguish between the dimensions on the test, which will result in smaller equating errors.
   b. Number of items on each subscale relative to the total number of items on the test – naturally, the more divergent the two forms of the test, the less equatable the tests.
   c. Different ability distributions of subpopulations on the dimensions. It was hypothesized that equating is more likely to fail with increased interaction of dimension and ability.
2. Conversely – when are the equating errors small enough to justify the same construct assumption?
3. As a practical implication of the research questions 1 and 2, one can ask whether a test can be scaled together to report a composite score (e.g. the ACT model) or should each subscale be considered separately (e.g. the SAT model) depending on the form and examinee characteristics?

The dissertation is structured as follows: first, (Chapter 2) Item Response Theory (IRT) models are described and multidimensionality and equating are introduced. This chapter also reviews some of the equating literature together with equating evaluation indices and literature on the interaction of equating and multidimensionality. Chapter 3 describes the methods used to answer the research questions posed above. Chapter 4 states the results obtained. Finally, Chapter 5 lists the shortcomings of the study and

proposes some further studies to both expand the current methodology, and further explore the general problem of unidimensional equating of multidimensional traits. Full result tables are included in Appendix A. Appendix B lists the SAS output code for the glm procedure used to parse the results. Appendix C illustrates sample sizes at the extremes of the population. Parscale code used to obtain the ability estimates is included in Appendix D.

2 Theoretical Framework

2.1 Item Response Theory (IRT) Models

Three models are commonly used in modern day psychometrics to describe examinee responses to individual dichotomous items. These are the One Parameter Logistic (1PL, a.k.a. the Rasch) model, Two Parameter Logistic (2PL) and the Three Parameter Logistic (3PL). The probability of a correct response on item i by an examinee with ability θ is given by:

$$P_i(\theta) = \frac{1}{1 + e^{-1.7a(\theta - b_i)}}$$

$$P_i(\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}}$$

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}}$$

for the 1PL, 2PL and 3PL respectively. In all of the above formulas *a* is the item discrimination, *b* is the item difficulty and *c* is the pseudo-guessing parameter. For the 1PL model the discrimination is assumed to be identical for every item on the test.

The polytomous items[2] are most often modeled by the Generalized Partial Credit model – the probability of an examinee with ability θ responding to the kth category of item i is given by:

$$P_{ik}(\theta) = \frac{\exp(1.7a_i \sum_{\upsilon=0}^{k}(\theta - (b_i - d_{i\upsilon})))}{\sum_{c=0}^{m_i}\exp(1.7a_i \sum_{\upsilon=0}^{c}(\theta - (b_i - d_{i\upsilon})))}$$

where m_i is the number of categories in the response to item j, a_i is the slope parameter, b_i is the item location parameter (characterizing the overall difficulty of the item) and $d_{i\upsilon}$ is the item i category k threshold parameter.

Software (e.g. Testfact (Wood et. al, 2003), Bilog (Zimowski, Muraki, Mislevy, & Bock, 1996), Parscale (Muraki & Bock, 1993)) is readily available to estimate item parameters for those models, and it's those parameters that are calibrated whenever IRT is used in equating.

2.2 Multidimensionality

It seems convenient to define multidimensionality as the lack of unidimensionality. Unidimensionality is intuitively simple – only one characteristic is measured by the test. Stout, (1990) illustrates the point quite well– "the dimensionality s of a test U is the minimal dimensionality required for Θ to produce a latent model (U, Θ) which is locally independent and monotonic." While there will always be nuisance variables measured – anxiety, motivation, etc, a cognitive test should be able to achieve

[2] i.e. items scored as wrong, (different degrees of) partially correct and correct.

Local Independence and Monotonicity[3] without those being included in the model. Thus, while the performance on the science test described in the introduction might also be influenced by the examinee's ability to read, follow directions, possibly reason logically, draw, and a number of other nuisance dimensions, it is hoped that the effects of those will be very limited.

One of the important statistics for assessing the degree of multidimensionality between tests is the correlation between traits. This is calculated with a Pearson correlation:

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

where σ_1 and σ_2 are the standard deviations of distributions of the abilities and σ_{12} is the covariance between the abilities.

The correlation is calculated either between examinee abilities (in a simulation study, since those are not observed), or raw scores on the subscales. While for a mathematics test the correlations between subscales are usually fairly high – around 0.95 for high school mathematics, some tests don't exhibit that high of a correlation. Generally, the higher the correlation, the more difficult it is to distinguish between constructs measured. It might therefore be possible to equate a purely algebra test to a purely geometry test without much detriment given the high correlation between those

---

[3] Local Independence means that conditioned on the construct the test is purported to measure, the person's responses to items are independent $P(U_1, U_2, \ldots U_n \mid \theta) = P(U_1 \mid \theta)\, P(U_2 \mid \theta) \ldots P(U_n \mid \theta)$ where U_i is the response of examinee with ability θ to item i.
Monotonicity means that the probability of endorsing an item increases with ability.

two abilities. While a purist might argue that in such a situation the same construct requirement is not fulfilled, this might be of little statistical or practical consequence provided all subgroups exhibit comparable performance on both subscales.

Item difficulty could be confused with multidimensionality (Ackerman, 1994) when one form subscale difficulty differs from the other form's subscale difficulty. This will render equating impossible, since unidimensional equating functions cannot account for dimension-specific changes in difficulty. The scores in this situation would never be equivalent across forms.

While today's commonly used IRT models are based on the assumption of unidimensionality, the presence of multidimensionality at the item level does not necessarily mean equating between forms is impossible. In fact, multiple dimensions can add to the predictor space and thus assess the (purported) one-dimensional construct better. Dorans (2004c) gives an example of "reading graphs and tables", which is a skill unto itself, but is also known to function differently for males and females. Still, it needs to be included in a social studies exam as part of the construct tested.

Ackerman (1996) developed a graphical representation of multidimensional tests which has proven useful in visualizing such tests. This representation is depicted in Figure 1 below.

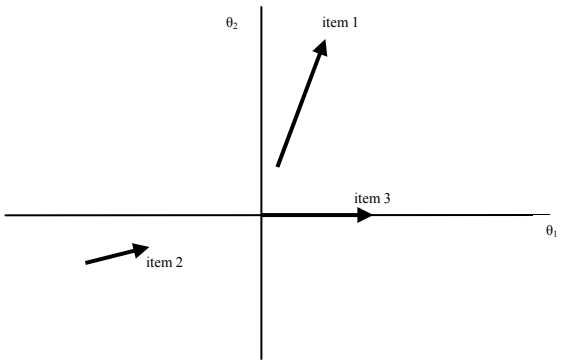

Figure 1: Graphical representation of multidimensional items (Ackerman, 1996)

Directional vectors are used to represent items. The length of a vector indicates the discrimination (item 1 is more discriminating than item 2); the angular direction of a vector indicates the composite trait that is being measured (item 3 measures θ_1 only, while item 1 measures θ_2 to a greater extent than it measures θ_1). The location of the vector indicates item difficulty (item 2 is less difficult than either item 1 or 3). Ackerman also suggests different arrowheads for different magnitudes of the 'c' parameter – open-tipped arrowhead for $c<0.1$, closed arrowhead for c between 0.1 and 0.2 and solid arrowhead for $c>0.2$. However, this unnecessarily complicates the picture and is not used here.

Notice in Figure 1 that item 3 measures only θ_1, while items 1 and 2 measure both θ_1 and θ_2 (with item 2 measuring θ_1 to a greater degree than θ_2; item 1 – the reverse). If a test consists only of items similar to item 3, it is said to exhibit approximate simple structure, if there are items like item 1 or item 2 in the test – it is called a nonsimple structure. Since in a practical testing situation those structures are more a matter of agreement than of any theoretical justification and implementation of nonsimple structure

to practical testing situations presents serious problems, a simple structure is most commonly assumed. This structure is going to be used in this study as well.

The situation described in the introduction in which the overall trait measured by the test shifts because individual items have been removed or added is schematically represented in Figure 2.

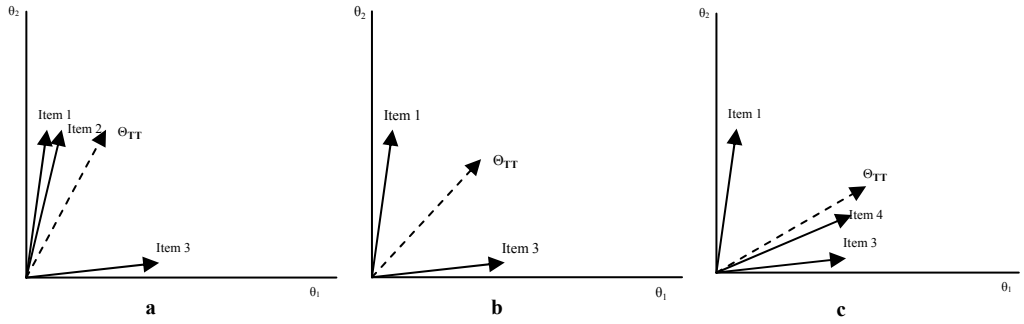

Figure 2: Graphical representation of the shifting Θ_{TT} caused by the shift in individual items.

In 2a) items 1 and 2 predominantly measure the ability θ_2, while item 3 predominantly measures ability θ_1. The overall test measures the direction indicated by the dashed arrow (called Θ_{TT} throughout this dissertation following Stout, 1996, or the reference composite). In 2b) item 2 has been removed; Θ_{TT} shifted more towards the θ_1 ability. In 2c) item 4 which measures θ_1 over θ_2 has been added to the test shifting the Θ_{TT} even further towards the θ_1. The question asked is if the same Θ_{TT} is measured a through c.

Currently, extensive research is being conducted on the multidimensional extensions of the IRT models (MIRT; e.g. Ackerman, 1994; Reckase, 1985, 1997; Reckase & McKinley, 1991). Most popular of those models are given by

$$P(X_i = 1 \mid \theta) = \frac{1}{1 + e^{-1.7(a^T\theta + d_i)}} \quad ,$$

$$P(X_i = 1 \mid \theta) = \frac{1}{1 + e^{-1.7(a_i^T\theta + d_i)}}$$

and

$$P(X_i = 1 \mid \theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7(a_i^T\theta + d_i)}}$$

for the 1PL, the 2PL and the 3PL respectively. Similar extensions exist for polytomous items (te Marvelde, Glas, van Landeghem & van Damme, 2006). As can be seen, the difference between the multidimensional and unidimensional IRT models described in section 2.1 is that both the discrimination and the ability parameters are now n-dimensional vectors (a_i^T in the equations above) representing the loading of the item discrimination and examinee ability on each of the n dimensions. d_i is a scalar related to the difficulty of the item. The above models are compensatory – an examinee can be very low on one ability and still endorse the item if she is sufficiently high on the other ability (or abilities). Noncompensatory IRT models have also been developed – generally those models have a multiplicative (rather than additive) ability. However, as they have not been widely used in research literature, they are not further considered here.

2.3 Equating

The role of equating is to make scores interchangeable across forms. It is mainly used to adjust for difficulty differences of forms. While it is always possible to "match up" the scores on one test to the scores of another test, this "matching up" has to fulfill very specific criteria (listed on the next page following Dorans & Holland, 2000) in order to be called equating; otherwise linking or concordance are performed. Prediction – a regression technique that predicts the scores on one test from the scores on the other, is used infrequently in cognitive testing. Before the linking is done a "logical evaluation of the similarity of the processes that produced the scores" (Dorans, 2000) needs to be performed and the strength of the relationship between scores needs to be evaluated.

Kolen and Brennan (2004, pp.7-8) list seven steps for implementing equating:

1. Decide on the purpose of equating;
2. Construct alternate forms;
3. Choose a design for data collection;
4. Implement data collection design;
5. Choose one or more operational definitions of equating – what types of relationships between forms are to be estimated?
6. Choose one or more statistical estimation methods;
7. Evaluate the results of equating.

Steps 1, 3, 5 and 7 are described in this section, steps 2,4 and 6 are described for this project in the methods chapter. Further description of step 7, specifically for this dissertation, is in Chapter 4.

Dorans and Holland (2000) identify 5 requirements for equating to be valid (following Lord, 1980).

1. Same construct requirement – forms need to be developed to the same specifications. This requirement is only seemingly trivial. In practice, it's very difficult to assess what the test measures. Dorans and Holland (2000) give an example of the (old) SAT verbal section which might be construed to measure verbal reasoning, "college-bound" vocabulary, white American domination, general intelligence, and so on, depending on the perspective (and agenda) of the person judging the construct. Even experts cannot reliably assess what the test measures nor, in a multidimensional case, partition the test into clusters measuring similar constructs. In a multidimensional case, the "same construct" requirement could also have multiple meanings. For instance, the consistency and quality of the dimensions measured by the test might need to be established. Moreover, individual items need to be classified according to what subscale they belong to (in a simple structure case) or to what extent they measure each subscale (in a nonsimple structure case). In fact, the constructs measured by a single item don't have to be cognitive at all for a test to be multidimensional – Kim and Lee (2006) for instance, have found that item format (multiple choice vs. constructed response) measured different competencies in the simulation study they conducted. The authors suggest a separate calibration be used for items depending on their format.

   There have been multiple studies trying to link the various sections of the (old) SAT to seemingly corresponding sections of the ACT. While those tests can be

described as measuring the same construct, linking has not always been successful and can never be called equating. If the tests are build to different specifications only concordance is possible – the scores in this case are not interchangeable.

2. Equal reliability requirement – both forms need to have the same reliability. For psychometric tests, reliability describes how well the observed score reflects the true score. The lower bound of reliability is called Cronbach's alpha and is given by:

$$\alpha = \frac{N}{N-1}\left[1 - \sum_j \frac{\sigma_j^2}{\sigma_Y^2}\right]$$

where N is the sample size, and σ_j and σ_Y are the standard deviations of the item and the test respectively.

3. Symmetry requirement – the function used to transform a score on form X to a score on the form Y scale has to be the inverse of the function used to transform a score on form Y to the score in the form X. This requirement eliminates simple linear regression (or any regression for that matter) as a form of equating, since $f(x) \neq f^{-1}(x)$.

4. Equity requirement – it is supposed to be a matter of indifference to the examinee which form she takes. This has the following immediate implications – firstly, the examinee should expect the same score regardless of which form of the test she took, and secondly – the distribution of scores given the ability should be the same on the original and equated forms. Lord (1980) defined it specifically as:

$G^*\left[eq_Y(x)\,|\,\tau\right] = G(y\,|\,\tau)\ \ \forall\tau$, where τ is the true score, eq_Y is an equating function used to convert scores on form X to form Y scale, G is the cumulative distribution of scores on form Y and G^* is the cumulative distribution of eq_Y. Examines with a given true score have identical observed score means, standard deviations and distributional shapes of converted scores. The conditional standard error of measurement at any true score should be equal on the two forms. Therefore, if one form measures some ability a bit more precisely than the other form, the equity requirement is not met (Kolen & Brennan, p. 11). However, under Lord's criterion equating is either not necessary, since the forms are identical, or is impossible, since any divergence between forms will never fulfill this requirement. This prompted Morris (1982) to suggest a first order equity property – examinees with a given true score have the same mean converted score on both forms:

$$E\left[eq_y(X)\,|\,\tau\right] = E(Y\,|\,\tau),\ \ \forall\tau$$

Similarly, the second order equity property – conditioned on the true score examinees have the same standard error of measurement on the two forms – is also commonly used to evaluate equating in simulation studies. Both first and second order equity properties are functions of unobservable true scores and therefore cannot be readily used in practice.

5. Population invariance requirement – the choice of subpopulations on which the equating function is calculated does not influence the conversion of scores from

one form to the other. There are two ways of looking at the equating functions for subpopulations. The first one is to apply the equating function calculated on the full population to subpopulation P_j while concurrently applying the equating function calculated on the subpopulation P_j (e_{pj}) to it, the other is to apply the e_{pj} to the full population P. In the first case, the e_p is treated as the "true" equating function and errors of the subpopulation-specific function are the errors due to equating, in the second case – the equating functions calculated on different subpopulations can be compared on the full population. The use of a function calculated on one subgroup on the other subgroup is "unlikely to be considered in practice", but "may be of interest from a research perspective" (Liu, Cahn & Dorans, 2006).

The requirements listed above are only theoretically straightforward. As mentioned previously, being able to assess what an item (and thus – a test) really measures can be a nontrivial endeavor. The reliability requirement is easily computed, but there has been no research to date on how close the reliabilities of two test forms need to be for the forms to be considered measuring the same construct. The population invariance is the most testable one, and can be used to evaluate the results of equating. If the population invariance assumption is violated there is strong evidence of an interaction between the form (presumably the difficulty of the form) and the group membership (i.e. the difficulty of both test forms differs across subgroups), which can result in differential item functioning of items or even of the whole form. Differential treatment (i.e. using

separate linking functions for subpopulations) has to occur if reproducing the distributions of scores on one form on the other form is the goal.

When linking fulfills the first four requirements, but is not population invariant it is called concordance. If concordance is implemented, linking functions should be obtained for important subpopulations (Dorans & Holland, 2000).

2.3.1 Data collection designs

There are several experimental designs commonly used when equating of two test forms is desired. Those are represented graphically in Figure 3 (adapted from Davey, Oshima & Lee, 1996) on the next page.

In the *random group design (equivalent groups design*) the examinees (from the same population) are randomly assigned to the form being administered. Any difference in score distributions in this design will be due to form, not to differences in examinee abilities (Han, Kolen & Pohlmann, 1997). Sometimes every other examinee is administered the same form – a procedure known as spiraling – which results in randomly equivalent populations receiving the two forms. Since each examinee takes only one form, it minimizes testing time, eliminates some of the practice and fatigue effects associated with taking a longer test, and is possible to administer in a high-stakes situation. However, large sample sizes are usually needed for this design. Additionally, it might be difficult to administer, if the forms consist of differently timed sections (Holland & Dorans, 2006). It is also impossible to administer forms in different test administrations, since ‘same population’ requirement is not met (unless the sample size is very large relative to the population size). In the *single group design* each examinee is

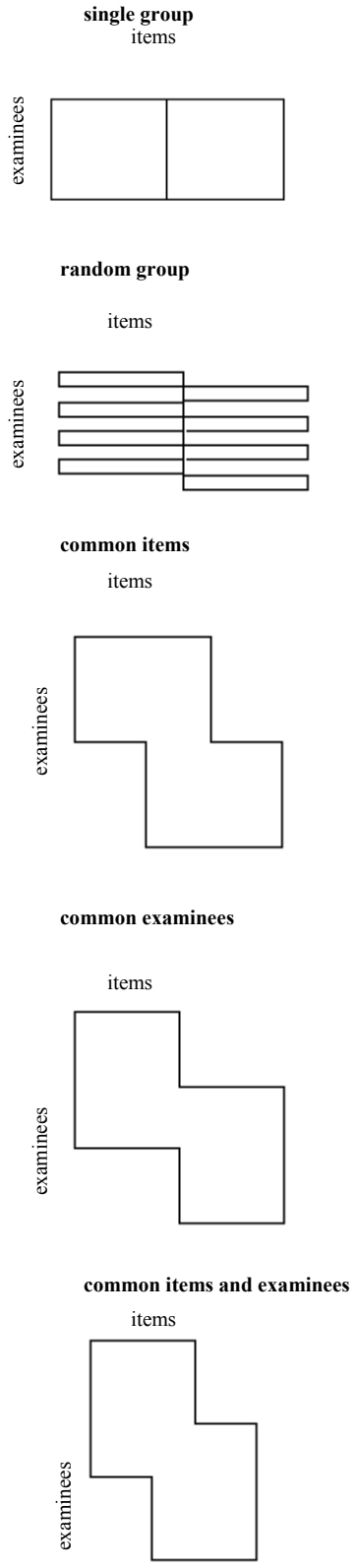

Figure 3: Data collection designs

given both test forms. In a real testing situation fatigue and practice effects are usually strong, and this design is not often used. In order to alleviate those effects, counterbalancing might be used – one form appears once in the first position (for the randomly chosen half of examinees), and once in the second position (for the remaining examinees). In practice, testing time has to be doubled to administer this equating design. However, a smaller sample size than for random group design is necessary if position effects are not present. Finally, the *common item nonequivalent group design* (NEAT design) is most often used in practical equating situations. In this design different populations of examinees are administered separate forms with some items in common between forms. The common items, called an *anchor,* should be representative of the test forms in terms of content and statistical characteristics, and should occupy a similar position within each form to control for position effects. A major advantage of this design is the chance to release non common items after they have been administered. The score on the anchor can either count towards the test score (internal anchor), or not (external anchor). Kolen and Brennan however suggest the common item nonequivalent groups design, while seemingly perfect, should be used with caution – strong statistical assumptions are necessary to separate the group and form differences. This separation gets more difficult with larger ability differences between the populations. The distinct populations might have to be combined (using weights) to form a synthetic population, but not all parameters of this synthetic population can be estimated (since the first population did not take the second form for instance). In this situation it is either assumed that the regression of each test form on the common items is the same for both populations, and that the conditional variance of each test form given the common items

are the same for both populations (the Tucker method), or that the true scores on form X and common items as well as on form Y and common items correlate perfectly for both populations and the measurement error variance is the same for each form on both populations (the Levine Observed score method).

Generally, the single group design is the best; however, because of administration constraints it's often impossible to implement in practice. Since this is a simulation study, with no administration limitations, single group design is going to be used. This, of course limits its applicability, although the results obtained here can be considered an upper bound for the possible equating errors. This issue is further discussed in Chapter 5.

2.3.2 Types of equating

The following methods are commonly used to equate scale scores or link true scores. Each of them has its own assumptions, advantages and disadvantages. The methods described below are described for the random and single group designs. Most of them have been adapted for the other equating designs (such as the common item equating), but those adaptations are reviewed only cursorily.

In the below descriptions X and Y are two forms of a test, as well as the scores on those forms. x and y represent a particular score, $\mu(x)$ and $\mu(y)$ denote the mean scores on form X and Y respectively and $\sigma(x)$ and $\sigma(y)$ denote the standard deviation of scores on those forms. e_Y is an equipercentile symmetric equating function used to convert scores on X to the scale of the score on Y. It is assumed that all forms fulfill criteria 1 and 2 of equating and all functions fulfill criteria 3, 4, and 5.

Mean equating

In mean equating the difference between the difficulty of two forms is considered constant across the whole ability scale. The equating sets the means of form X and Y to be equal and scores that are an equal distance away from their means are set equal.

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

Linear equating

In linear equating not only the means, but also the standard deviations of the scores are matched; the difference in difficulty of the forms is allowed to vary across the score scale. The standardized deviation scores (z-scores) are set to be equal.

$$\frac{x - \mu(X)}{\sigma(X)} = \frac{y - \mu(Y)}{\sigma(Y)}$$

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

With the linear equating, as with mean, equating it is possible for the equated score to be above or below the possible range of raw scores (i.e. negative for higher than the maximum raw score). Kolen and Brennan (2004, p.34) suggest 2 methods of adjusting for this. One is to allow the maximum and minimum equated score to float and the other is to truncate all negative scores to zero and set all scores above the maximum raw score to the maximum raw score. They note that sometimes the issue of choosing the

adjusting method is of "no consequence" since no one achieves the extreme scores. Often specific tests will have policies as to the assignment of the highest possible raw score. The SAT program for instance assigns the highest possible raw score to the highest possible scale score no matter what the equating results might indicate.

Equipercentile equating

Equipercentile equating uses a curve, rather than a line to describe differences in difficulty between forms and matches all moments of the form X score distribution. Angoff's (1971, p. 563) definition of equipercentile equating states that "true scores, one on form X and the other on form Y (where X and Y measure the same construct with the same degree of reliability) may be considered equivalent if their corresponding percentile ranks in any given group are equal".

Braun and Holland (1982) restate this definition: "The function e_Y is the equipercentile equating function in the population if when the cumulative distribution function of scores on form X converted to the form Y scale is the same as the c.d.f of the scores on form Y":

$$e_Y(x) = G^{-1}[F(x)]$$

where F is the cumulative distribution function of X, G is the cumulative distribution function of Y. This definition of equipercentile equating is also used in "Test equating, scaling and linking" (Kolen & Brennan, 2004).

In practice, the discreteness of the scores might cause the distributions of the two scores to differ. However, in a large-scale testing situation "not being able to achieve the equal distribution goal is often more a theoretical consideration than a practical one" (Kolen & Brennan, p. 48).

Equipercentile equating is straightforward to implement, has few assumptions (particularly regarding model fit) and is used in multiple testing programs (e.g. the SAT, which also uses linear equating whenever appropriate) It also does not have the problem of linear and mean equatings where the form Y equated scores are outside the possible range of form X raw scores.

The algorithm for finding the equipercentile equivalent has 3 steps:

1. For a score x on form X find the percentage of examinees at or below this score.
2. Find a score y on form Y that has the same percentage of examinees at or below it.
3. Score x and y are considered equivalent.

In the case where the score distributions are identical in shape and differ only in their means and standard deviations, linear and equipercentile equating are equivalent. Linear equating introduces less random error than equipercentile equating, and it has been shown that equipercentile equating is linear equating with some variability added (von Davier, Holland & Thayer, 2004).

IRT true score equating

The equating method described below assumes the items follow one of the Item Response Theory (IRT) models described in section 2.1 IRT true-score equating links true (i.e. unobserved) scores between test administrations.

True score is the sum of *p(θ)*s for all items on the test:

$$\tau_X(\theta_i) = \sum_{j:X} p_{ij}(\theta_i; a_j, b_j, c_j)$$

where *i* denotes the examinee and *j* an item. IRT true score equating sets the item parameters to be equal between the two distributions. True scores on form X and Y are equivalent for a given θ. This equating method is very convenient with the common item nonequivalent group designs and, assuming the IRT model holds, is always population invariant (Han, Kolen & Pohlmann, 1997). However, observed, rather than true scores are present in testing situations, and there have been "philosophical differences" (Harris & Crouse, 1993) as to whether true score equating should be done in practice.

In the data collection designs other than the single group design, item parameters have to be calibrated i.e. put on the same scale for the two forms. Methods of estimating item parameters have been a subject of research. Concurrent estimation (all item parameters from all forms are estimated together) is the most popular one; however, separate estimation (item parameters are estimated separately for each form then transformed to the same scale) has also been researched, as well as the fixed parameter (anchoring) estimation. In the separate estimation method matching item parameter means or means and standard deviations are used. The mean/sigma method uses the

means and standard deviations of the 'b' parameter estimates from the common items to put all the parameters on the same scale. The mean/mean method uses the means of both 'a' and 'b' parameters for the same purpose. Item parameter scaling is only needed when the groups taking the two forms are nonequivalent, but it can reduce the estimation error by reducing the differences in the θ scale due to sampling error. There are also characteristic curve methods (such as the Stocking-Lord and Haebara) where all parameters for all items are considered simultaneously. Those usually produce less error (Kim & Lee, 2006) than the moments methods. They involve, however, extremely computationally intensive iterative procedures and are not further discussed here. Discussion, description and criticism of those methods are considered in detail in Kolen and Brennan (2004, p. 168). Davey, Oshima and Lee (1996) theoretically extended and adapted the unidimenisional linking procedures to a multidimensional case. Those extensions (called direct method, equated function method, test characteristic method and item characteristic function method) were tested in a simulation study by Oshima, Davey and Lee (2000).

In research comparing the estimation methods it was generally found that for large sample sizes (around 3000 examinees per form) concurrent estimation had lower errors (Hanson & Beguin, 2002). When the number of common items is small, the separate estimation is better (Kim & Cohen, 1998). This last study concluded that further research was needed to reaffirm this estimation method's superiority over separate estimation. If fact, the authors go on to say that using separate estimation and obtaining two sets of parameters might help pinpoint potential problems in scaling. Concurrent

estimation might even be better than the anchoring (fixed parameter) estimation, since larger sample size is used for parameter estimation.

How true-score equating works in practice is represented graphically in Figure 4 below where test characteristic curves for 2 forms are depicted.

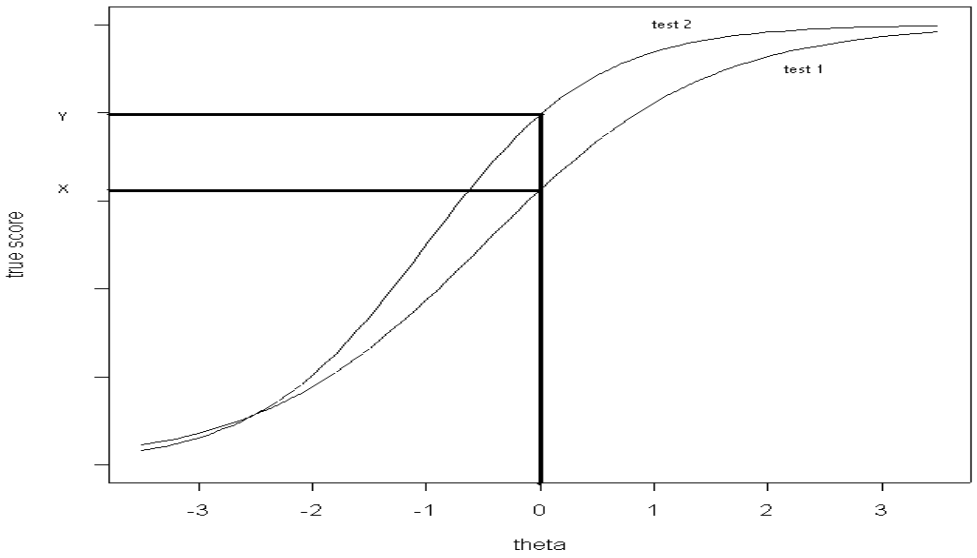

Figure 4: Illustration of true-score equating.

For a given value of θ (0 in the figure above) the form 1 true score is x, while form 2 true score is y. Assuming the item parameters for those forms are on the same scale, x and y are considered equivalent.

Notice that technically, IRT true-score equating is not equating according to the Dorans and Holland definition described above. While the population invariance requirement is theoretically fulfilled, it has never been tested in practice. The symmetry requirement might also be called into question because of inherent unavailability of true scores. However, since it is the "most commonly used" form of linking it's going to be used here as well.

2.3.3 Some studies comparing equating methods.

Care should be employed when using the model-based (IRT) equating methods in real testing situations since they have strong distributional assumptions for parameters. The assumption of normality for instance, is hard to keep for smaller sample sizes. Additionally, the parameter estimates are frequently not stable (with the 'b' parameter generally considered more stable than the 'a' parameter), and there usually is a large amount of measurement error on the θ values, especially at the extremes of the scale. This is especially pertinent if the 3PL model is used since there will be no relationship below the sum of the pseudo guessing parameters.

Equipercentile and IRT true-score equating outperform other methods when the relationship is not linear, i.e. the shapes of score distributions (3rd and 4th moments) differ (Ree, Carretta & Earles, 2003) and accuracy is required along the whole score scale. In the random groups design, linear equating has been shown to have less random error for normally distributed scores than equipercentile equating, and thus requires smaller sample sizes to achieve the same equating precision (Kolen & Brennan, 2004 p. 256). Similarly, single group linear equating usually has less error than the random group linear equating. In the single group design systematic error is more likely to occur with the need to counterbalance practice and fatigue effects (Kolen & Brennan, 2004, p. 15).

Many studies have compared equating methods for unidimenisional tests. Han, Kolen and Pohlmann (1997) used the ACT datasets to compare IRT true and observed score equating with equipercentile equating. IRT true-score equating had more stable equating results than the other two equating methods. IRT true-score equating is sample independent. Since the ACT uses a randomly equivalent group design, any difference in

score distributions is due to form, not to examinee. The authors suggest equating the test onto itself (either by using the entire group or split-half groups) to assess the stability of the equating results. In the study, the IRT true score equating had the most minimum means of loss index, so it's more stable than equipercentile equating. The authors also selected Math and English forms by difficulty (least, median and most difficult). The most difficult forms were equated to the other two forms and the median form was equated to the least difficult form. The equating differences were largest when the difficulty of the forms was most different (the result was not as clear for the mathematics forms).

Tong and Kolen (2005) used the Iowa Test of Basic Skills (also a random group design) to compare the performance of the equipercentile, IRT true score and IRT observed score equating using three criteria – same distribution property, first-order equity property and second-order equity property[4]. The authors concluded that if the raw score distributions on the forms are similar all three methods lead to good equating. When raw score distributions are dissimilar, IRT true score method performs best in preserving the first order equity property while equipercentile and IRT observed score method preserve the second order equity property best. Equipercentile and IRT observed score method also work well if the same distributions are to be preserved. The greater the difficulty difference between the tests the more difficult it is for the first and second order properties to hold.

The authors used the nonparametric Kolmogorov t-statistic which looks at the largest difference between the two relative cumulative distribution functions to test the

[4] As a reminder – the first order equity property means that conditioned on the true score, examinees have the same scale score. Second order equity property – conditioned on the true score examinees have the same standard error of measurement on the two forms.

equal distribution property. The smaller the difference, the more similar the two distributions. This t-statistic has an advantage over the more popular methods described below of following a theoretical distribution. As the difference in CDFs increased, the first order equity property statistic did not increase, but the second order equity property statistic did. This increase was more pronounced for the IRT true score method. This was not entirely consistent with the hypothesis that as the difference in scores increases the first order equity is not preserved. The authors also used the standardized difference to look at the results (standardized in order to make the results for different forms comparable). The capitalization subtest had zero scale score differences, but not the maps and diagrams (which had the highest difference between the means on the two forms). Second order equity property is less likely to be preserved with a high difference in raw score distributions.

Tong and Kolen concluded that all three equating methods lead to "reasonably similar" scale score distributions, with IRT true-score equating performing worse than the equipercentile equating on matching the scale score distributions (since it matches the estimated true scores).

Tong and Kolen also conducted a simulation study with increased "b" parameters – adding 0.2, 0.4, 0.6, 0.8, 1 and 1.2 with the IRT true and observed score equating. The observed score equating had much smaller t-values for b=0.4 and above added, so it preserves the distributions better.

The authors suggest that if tests differ substantially in their raw score distributions, use equipercentile and observed score to preserve the distributions and the

true-score to preserve the first order equity and observed score to preserve the second order equity.

2.3.4 Evaluating equating results

When can the tests be considered equated? How do we know if equating worked? Evaluating the results of equating is a necessary final step of the procedure. However, no universal procedure has been established so far for this purpose. What follows is a review of commonly used procedures.

How do we know if we should even attempt to link forms? Dorans (2000) introduces the coefficient of alienation: $\sqrt{(1-r^2)}$ where r is the correlation between forms. The coefficient of alienation is a measure of statistical uncertainty that remains after inclusion of information from the predictor variable. In equating, the predictor variable is the form to be equated. Dorans posits that "if a predictor cannot reduce the uncertainty by at least 50 percent it is unlikely that it can serve as a valid surrogate for the score." By simple algebra it follows that only correlations above 0.866 reduce the uncertainty by more than 50%.

Some of the evaluation measures are definitional in nature and therefore, at least conceptually, if not computationally, straightforward – mean, linear and equipercentile equating methods match the score characteristics of the distributions (first moment, first and second moments and all moments respectively). All equating by definition has to preserve equity (ideally the Lord's definition of it, but in practice – preservation of the weak equity is checked), but equity is often difficult to compute and explain (Harris &

Crouse, 1993). There has been some research (Bolt, 1999) on the degree to which the equity fails to hold.

When true abilities are known, as is the case in simulation studies, one might look at the misclassification rates of equating. Fitzpatrick and Yen (2001) looked at the false negative classification[5] rates at the 25th percentile, arguing that this is a standard pass percentile for high-stakes state tests. Naturally, one does not want to classify a student as failing if the student's ability is high enough to pass the test. (A mistake in the opposite direction, while carrying the same statistical importance, does not have grim consequences for the examinee). Similarly, one could look at what Ree, Carretta and Earles (2003) call the impact analysis – after equating, the test forms should have not only the same proportion of examinees passing the test, but also the proportions of examinees who passed one test form and failed the other should be the same. The authors suggest that the impact analysis be performed on the whole population and gender and race/ethnicity groups to evaluate the results of equating.

Equating a test to itself has been suggested by some researchers (Han, Kolen & Pohlman, 1997; Harris & Crouse, 1993). It's convenient, especially since the 'true' state is known (as is also the case when using generated data), but no equating will always be the best solution. Additionally, results obtained might depend on the starting form (Harris & Crouse, 1993). Simulated data problems will favor the equating method that uses the way the data were simulated (IRT equating when IRT models were used for instance). No equating is also a good solution if the distributions of scores on the two forms differ only by random error (as confirmed by the chi-square test for equal distributions). In this case equating would introduce only more random error.

---

[5] Passing examinees classified as failing.

In order to evaluate the results of IRT true-score equating, the differences between equated scores and the criterion scores are considered.

Han, Kolen and Pohlmann (1997) used weighted and unweighted mean signed difference (bias):

$$MSD_u = \frac{\sum_{i=0}^{k}(T_i - O_i)}{k} \qquad MSD_w = \frac{\sum_{i=0}^{k} f_i(T_i - O_i)}{N}$$

mean absolute difference

$$MAL_u = \frac{\sum_{i=0}^{k}|T_i - O_i|}{k} \qquad MAL_w = \frac{\sum_{i=0}^{k} f_i|T_i - O_i|}{N}$$

and root mean square difference

$$RMSD_u = \sqrt{\frac{\sum_{i=0}^{k}(T_i - O_i)^2}{k}} \qquad RMSD_w = \sqrt{\frac{\sum_{i=0}^{k} f_i(T_i - O_i)^2}{N}}$$

to evaluate the results of their equatings. In all of the above formulas T_i is the equivalent score, O_i is the criterion score for the T_i, k is the number of score points (for dichotomous forms – the number of items on the test); f_i is the frequency of the equivalent score O_i and $\sum(f_i)$ is equal to the total sample size N. For IRT true-score equating T_i and O_i are the two

true scores on the two forms. Those equating criteria are well established in literature (Livingston, Dorans & Wright, 1990; Harris & Crouse, 1993).

The disadvantage of using bias to evaluate the results of equating is that it will cancel out, unless it's in one direction across the whole of the score range, however, it can be "helpful diagnostically" (Livingston et. al, 1990) to investigate why the values of other indices are high. As such, it will be used in this study as well.

Using weights (in mean absolute difference) allows for points that occur frequently to be given more emphasis. No weights might be used when the whole score scale is considered. Harris and Crouse warn against using weights, especially when equating data from a pilot study for instance – examinees in the operational assessment might score at the extremes at higher rates than during the pilot.

The magnitude of the indices is used to evaluate the results, but what amount of difference is significant in equating has not been determined. Livingston et al (1990) set a cutoff for RMSD saying that a value of 5 or greater might imply problems in equating. Bias will lead to some values canceling out, but Livingston et al (1990) said that it can be "helpful diagnostically" to investigate why RMSD value is high.

For the multidimensional case, if the true ability (or rather – abilities) is known (i.e. simulated) Bolt (1999) introduces the conditional bias of equating.

$$d_l(\theta) = E_x[X \mid \theta] - E_y[x(Y) \mid \theta]$$

Where $\theta = (\theta_1, \theta_2)$ is the two-dimensional ability vector. This gives an indication of how well the equating function has matched expected scores. A positive value for $d_l(\theta)$

indicates that the expected performance at a given ability is greater on test X than test Y as a result of applying an equating transformation. Bolt suggests $d_l(\theta)$ might also serve as the local index of equity performance measuring how equating bias changes as a function of θ. By integrating over all θ, weighted average bias is introduced:

$$wad_l = \int_{\theta_1}\int_{\theta_2} d_l(\theta) f(\theta) d\theta$$

Where f(θ) is the bivariate density function of theta. Since wad is weighted by f(θ) the size of the population influences the size of the bias.

Finally, Bolt combines the first and second order equity criteria in the total conditional variance (tcv) measure:

$$\begin{aligned} tcv(\theta) &= E_Y\left[x(Y) = E_X(X) \mid \theta\right]^2 \\ &= d_1^2(\theta) + Var_Y[x(Y) \mid \theta] \end{aligned}$$

tcv gives a measure of how well the equating transformation predicts examinee expected score on X given the score on test Y. By adding $Var_X(X \mid \theta)$ to tcv, Bolt obtained an index measuring how well the equating fulfills Lord's definition of equity. This index,

$$\begin{aligned} d_{1.2}(\theta) &= E_Y\left[x(Y) - E_X(X) \mid \theta\right]^2 - E_X\left[X - E_X(X) \mid \theta\right]^2 \\ &= tcv(\theta) - Var_X(X \mid \theta) \end{aligned}$$

also assesses the accuracy of the expected score on X by Y compared to the actual score of X.

Harris and Crouse (1993) concluded their review of equating results by saying that all of the methods are valuable, but more research needs to be done. They admit that "it appears likely that the summary indices will continue to be used on the basis of their prominence in literature", not necessarily on the basis of their applicability to a particular equating situation.

2.4 Intersection of the equating and multidimensionality

To date, there has been very little research on the impact of multidimensionality on equating. Studies have generally concluded that the impact of multidimensionality on equating, at least the accuracy of the IRT true score equating, appears minimal. Most researchers resort to citing Wang (1985) who claimed that it's negligible as long as the same linear composite of latent traits (reference composite) underlies the item responses on both tests, but little thought is given to the degree of "sameness" that has to be achieved.

While there are multidimensional scaling methods being developed (Davey, Oshima & Lee, 1996), it is more common to use unidimensional equating for multidimensional tests. The rationale for it is two-fold. Firstly, on a test designed to be unidimensional, it is hoped that the degree of multidimensionality is not going to be severe enough to impact the quality of equated scores. If the test is designed as multidimensional, simple structure is assumed and each dimension is scaled separately (as is the case with NAEP). Complicating the matter further, multidimensional equating

methods are currently at the theory stage of development and have not been used in any operational test.

The studies concerned with equating in the presence of dimensionality assume construct equivalence. Usually, the same mix of items is used between forms. Most of the literature so far has considered the performance of IRT true-score equating under multidimensionality.

Camilli, Wang and Fesq (1995) found that the IRT true-score equating of the LSAT is relatively robust to multidimensionality. LSAT, through factor analysis, was found to measure two reasoning abilities. The authors divided the whole test into a homogeneous subtest (consisting of all the items) and a heterogeneous subtest (consisting of the items measuring the two abilities separately). Each of those was calibrated separately and IRT true-score equating was conducted for the two parts. The differences in the true score tables between the homogeneous and heterogeneous subtests were compared. Since those differences were small, the authors concluded that the true-score equating is robust to multidimensionality at least for the LSAT case (correlation of 0.7 between the factors).

De Champlain (1996) considered a practical possibility in which the dimensionality of a test differs depending on the racial/ethnic group of examinee. He discovered that a 3 dimensional model fit the Hispanic LSAT test takers better than the 2 dimensional model (which fit the black and white test takers). The author tried to test the population invariance of true-score equating by obtaining separate 3PL estimates for three racial groups then scaling them to one of the subscales using preoperational form

parameters using the characteristic curve method. Estimated true scores were obtained and transformed to the LSAT score scale by using CSEM DIFF (Dorans, 1984):

$$CSEM\ \ DIFF(\hat{\theta}_i) - \left(\left(\sum_{i=1}^{n} P_i(\hat{\theta}_i) * Q_i(\hat{\theta}_i) \right) * A^2 \right) * \sqrt{2}$$

Where $Q_i(\hat{\theta}_i) = 1 - P_i(\hat{\theta}_i)$ and A is the slope transformation that places raw scores on the LSAT reported scale.

The differences in means were small – ranging from 0.01 to 0.4 for Hispanic examinees (difference between the mean score obtained using just the Hispanic population equating function and either the Caucasian population equating function or the whole population) and between 0.02 and 0.26 for black test takers. The author concluded that the differences were negligible throughout the entire ability scale, even if they were larger at the low end of the scale, therefore the equating function obtained from all examinees does not penalize the minority examinees.

One study to date has examined the performance of traditional equating (either mean, linear or equipercentile) under multidimensionality – Bolt (1999). He compared the equipercentile, linear and true-score IRT equating for a two-dimensional test and found that IRT true score equating performs well for examinees high on both abilities. Equipercentile equating performs poorly for those abilities, but is considerably better then either the linear or IRT method for examinees low on both abilities. Equipercentile equating performed just as well as the IRT true-score equating on the error-measurement indices (described in the previous section) and better for low (0.5 and 0.3) correlations.

Surprisingly, the linear method does not appear to be affected by the increase in correlations between θs. Bolt noticed that even in the unidimensional case the equity criteria were not well satisfied (this, he suggested, might indicate the shortcoming of the indices he used). IRT true-score equating performed as well with high correlations between θs as the equipercentile and linear equating and slightly worse than the equipercentile method for lower (<0.5) correlations. Bolt concludes that "none of the methods is superior to the others for all examinees."

While Davey, Oshima and Lee (1996) claim that it is not possible to link tests with randomly equivalent groups of examinees and different tests since no method developed to date can link tests with a different number of underlying dimensions, Bolt explicitly suggests studying the effect of adding another dimension to the second form that was not present in the first form. He gives an example of a reading assessment which can have dimensions connected to the passage effects that the previous administration did not have. Some of the simulation settings (described in the next chapter) are going to explore this situation.

3 Methods

As can be seen from the literature review above, there has been very little research on the effects of multidimensionality on equating. Moreover, the research extant treats multidimensionality as consistent form-to-form and evaluates the results of equating within that framework. Compatibility of constructs is also always assumed in those studies. This dissertation attempted to reverse this thinking by using equating as a tool to evaluate the effect of varying degree of multidimensionality.

Errors in equating two two-dimensional test forms differing on their Θ_{TT} were examined in a simulation study.

3.1 Data generation

Responses of 4000 examinees to forty dichotomous items on two forms were generated (80 items total). Two IRT models described in section 2.2 were used – the M3PL (Reckase & McKinley, 1991) and the multivariate Rasch model (Reckase, 1985) The Rasch model was used in the hope of removing the possible effect of varying item discrimination on the results. Additionally, this model is commonly used in state assessments and is thus considered of practical importance. Item parameters were simulated to follow the distributions outlined by Donoghue and Allen (1993) – the discrimination ("a") parameter followed a lognormal distribution with mean zero and standard deviation 0.35, the difficulty ("b") parameter followed a standard normal distribution and the guessing ("c") parameter was set to 0.2. For the Rasch model only the "b" parameter was generated from the distribution above; the "a" parameter is commonly set to 1.

Correlations between dimensions were set to 0.5, 0.7, 0.9 and 0.95. Since one of the assumptions of IRT true-score equating is that of unidimensionality, it is felt that for correlations lower than 0.5 the errors of equating would have been large due to the severe multidimensionality, regardless of the other variables investigated here.

Three conditions of examinee ability distribution were investigated. All examinees were simulated to have a standard normal distribution on the second dimension. Half of the examinees were simulated to have a standard normal distribution of ability on the first dimension. The other half of the examinees had the mean ability increased by either 0.5 or 1 (corresponding to either half or full standard deviation) respectively. Those setting are depicted in Table 1 (for dimension 1 only) for easy reference. Those adjustments were applied to both forms.

Table 1: Mean ability of the examinees on the first dimension.

Dimension 1 (θ_1)	
50% of the population	**50% of the population**
0	0
0	0.5
0	1

Note that the dimensions are not independent (they are connected by the correlation) and thus two bivariate normal distribution will be simulated (one for each half of the population) and then combined to form a full population for estimation.

Single group design was used – same abilities were used to simulate item responses to items on both forms, and no increase in ability form-to-form is assumed.

Simple structure was used; number of items was distributed among the two forms according to the Table 2 below.

Table 2: Distribution of the number of items on each form, each dimension.

Form 1		Form 2	
1st dimension	2nd dimension	1st dimension	2nd dimension
20	20	20	20
20	20	21	19
20	20	25	15
20	20	40	0
0	40	40	0

The situation depicted in the first row of Table 2 provided a baseline for the equating errors. This is especially important at lower correlations between dimensions where the equating errors and correlation interact (per violations of the assumptions of IRT true-score equating) regardless of the shift in Θ_{TT} (which this study was trying to capture). Note that from the measurement perspective, it is not known how many items (and at what correlations) are necessary to be able to identify a cluster (Reckase, 1997). Zhang (2004) for instance, in his description of the dimensionality assessment program DETECT suggests that the users should disregard clusters with fewer than 5 items, since this number does not provide a stable solution to the partitioning algorithm. Therefore, the 20-20:35-5 split was not examined since it might not be differentiable from the 20-20:40-0 setting, especially at higher correlations.

Standard dichotomous item response generation procedures were used. This is a fully crossed design with 2*4*3*5=120 settings (models*correlations*abilities*form differences). The calculated results are based on ten simulation runs; the estimated results

are based on 50 simulation runs. All programming was done in R. For ease of programming, the data was generated separately for each form.

3.2 Equating

For each of the simulation settings and replications the two forms simulated were equated using IRT true-score equating as described in section 2.3.2. Item and ability parameters were obtained in separate calibrations using Parscale. Since single group design was used, it was not necessary to calibrate the item parameters – they are assumed to be on the same scale (Kolen & Brennan, p. 166). Form 2 was equated to form 1.

Preliminary results indicated that the unidimensional ability retrieved by Parscale most closely approximates the average of the two-dimensional abilities weighted by the number of items on each ability[6]. Therefore, equating was done along this line (i.e. the Θ_{TT} line). In order to keep the variances of the calculated ability consistent with the variances of the simulated abilities, this weighted average was divided by the cross product of the weights – $a^2+2ab\rho+b^2$, where a is the number of items at trait 1 divided by the total number of items, b is the number of items at trait 2 divided by the total number of items and ρ is the simulated correlation between trait 1 and trait 2. Notice that the function mapping a (θ_1, θ_2) pair to Θ_{TT} in this case is not 1-1, i.e. more than one pair of multidimensional abilities results in a given θ_{TT}. For example, a θ_{TT} of 1 on a test with 10 items from one dimension and 20 items on the other dimension could obtained by, among others, pairs (3,0), (3/2, 3/4), (1,1), (-20, 11.5), etc, all of which are solutions to the equation

---

[6] This statement is correct for simple structure and non-aberrant item discrimination parameters. Other combinations might be better approximated by the theoretical θ_α index of Zhang & Stout (1999).

$$\tfrac{1}{3}\theta_1 + \tfrac{2}{3}\theta_2 = 1$$

Note, that while (-20, 11.5) pair is a rather counterintuitive (θ_1, θ_2) combination, especially for nonzero correlations between normal traits; it is unlikely, but not impossible for an infinite population (which, in effect, is what a Monte Carlo study simulates). The equation above has an infinite number of solutions – those solutions can be represented graphically (in two dimensions) as lines perpendicular to the Θ_{TT} line at θ_{TT} – the dashed lines in Figure 5 below:

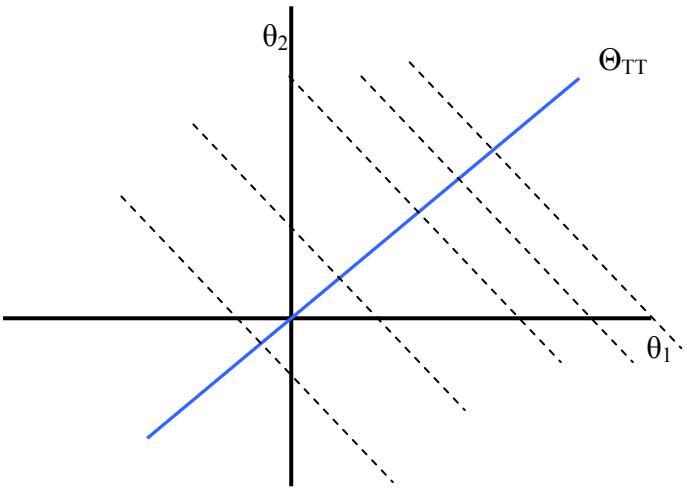

Figure 5: The (θ_1, θ_2) equivalence on Θ_{TT}

If the number of items at each of the abilities is known, for any (θ_1, θ_2) we can find the corresponding unidimensional ability (θ_{TT}). Unidimensional equating can be conducted using this θ_{TT} by finding the true scores on each of the forms to be equated. This was discussed in detail in section 2.3.2 and is illustrated again in Figure 6 below, where true score B is considered equated to the true score E.

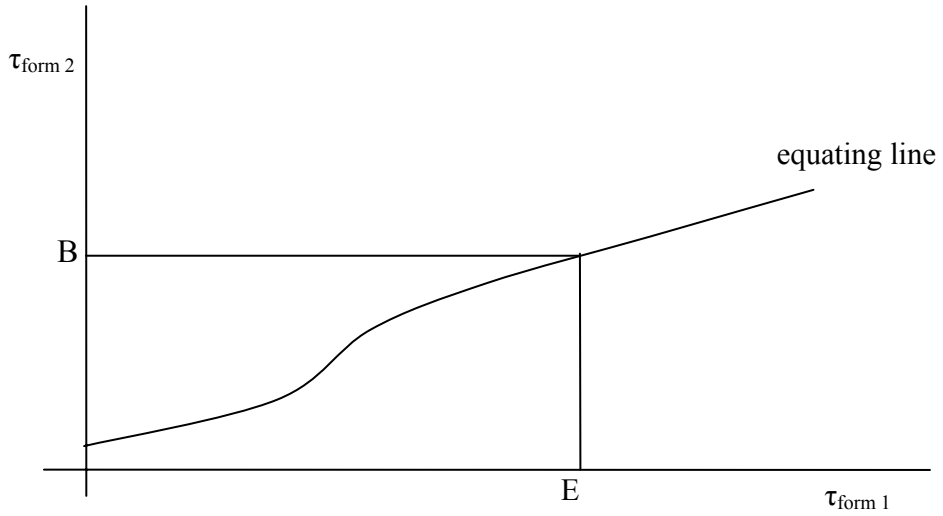

Figure 6: True score on form 1 ($\tau_{form\ 1}$) to true score on form 2 ($\tau_{form\ 2}$) equating line.

However, this equating line is along the unidimensional substitute for the multidimensional ability (later called calculated θ). In reality, for a two dimensional situation, there are two abilities contributing to θ_{TT}. A different viewpoint of this equating with an explanation how to find point B is depicted in Figures 7a through 7e on the following pages.

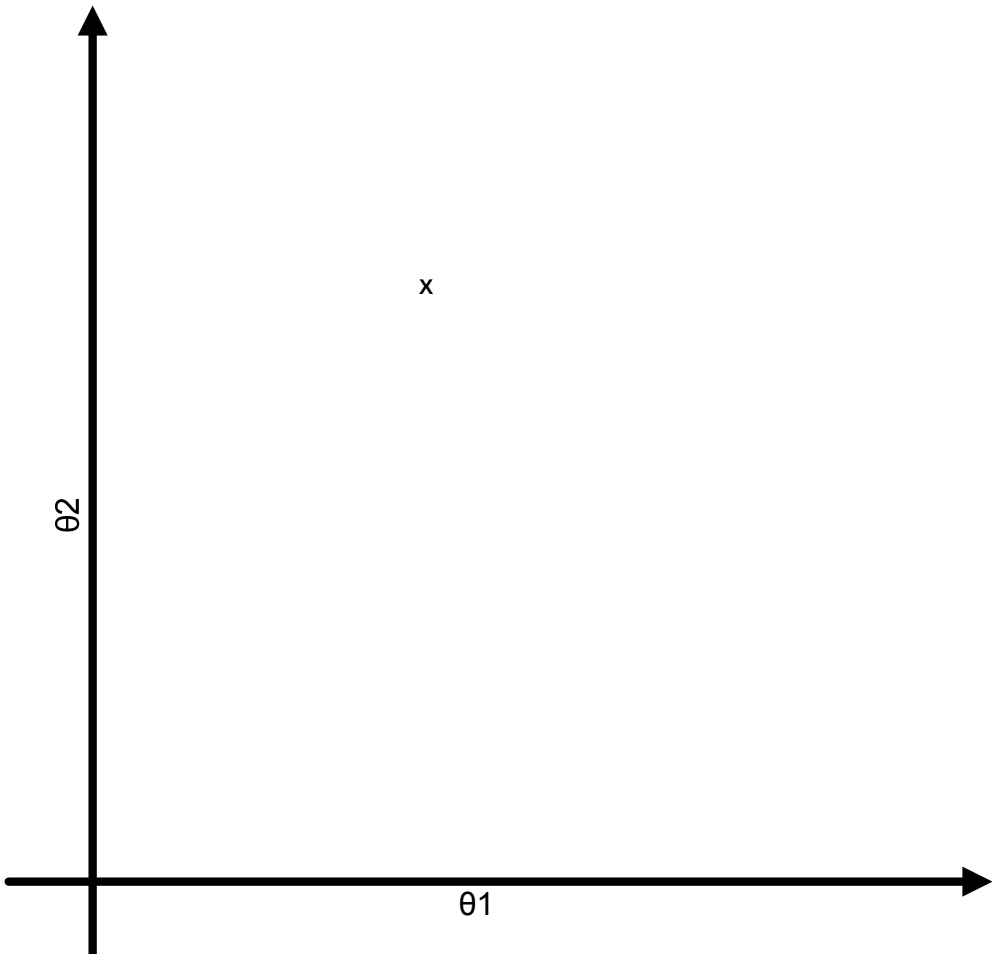

Figure 7a: Unidimensional equating in the multidimensional $\theta_1\theta_2$ plane

In Figure 7a above, x represents an examinee of the two-dimensional ability (θ_1,θ_2).

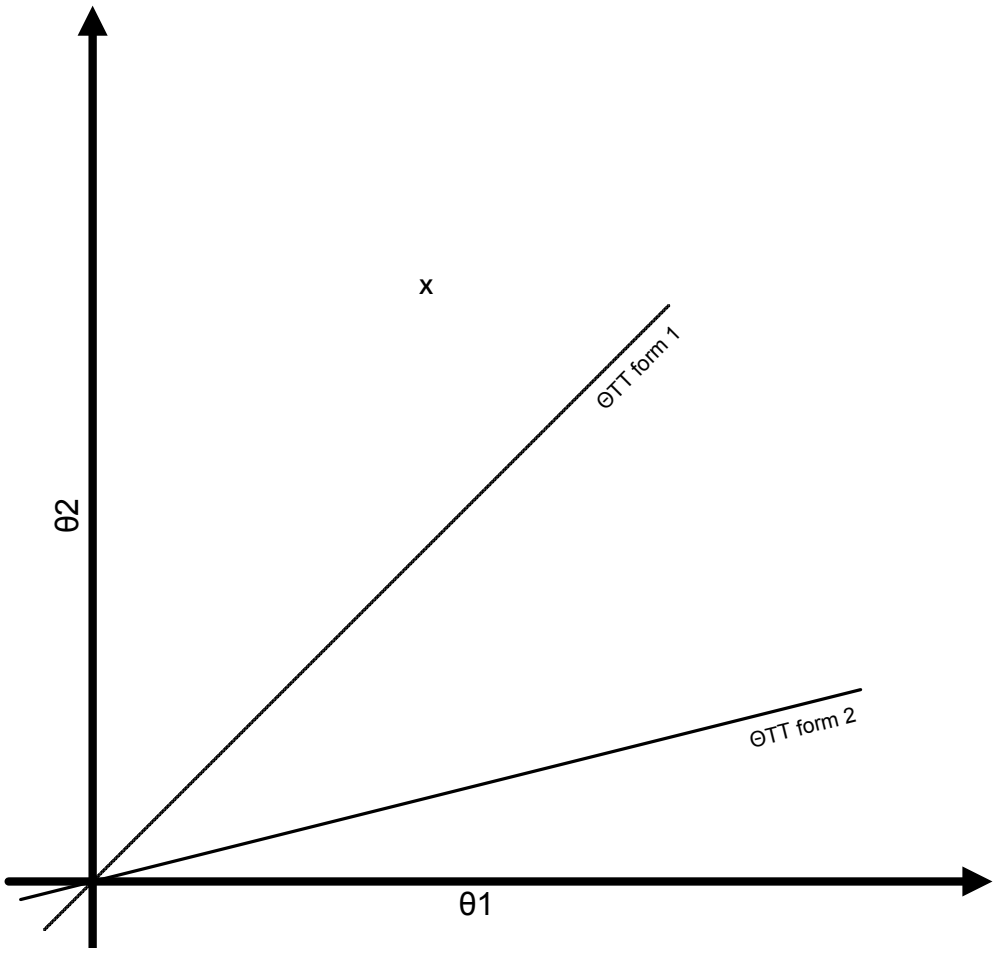

Figure 7b: Unidimensional equating in the multidimensional $\theta_1\theta_2$ plane

In Figure 7b, the two lines indicate the unidimensional composite ability (Θ_{TT}) measured by each of the test forms – form 1 contains the same number of items sensitive to θ_1 as to θ_2 (the line is at a 45 degree angle to each of the ability lines), form 2 has more items sensitive to the θ_1 dimension (the line is closer to the θ_1 axis than to the θ_2 axis).

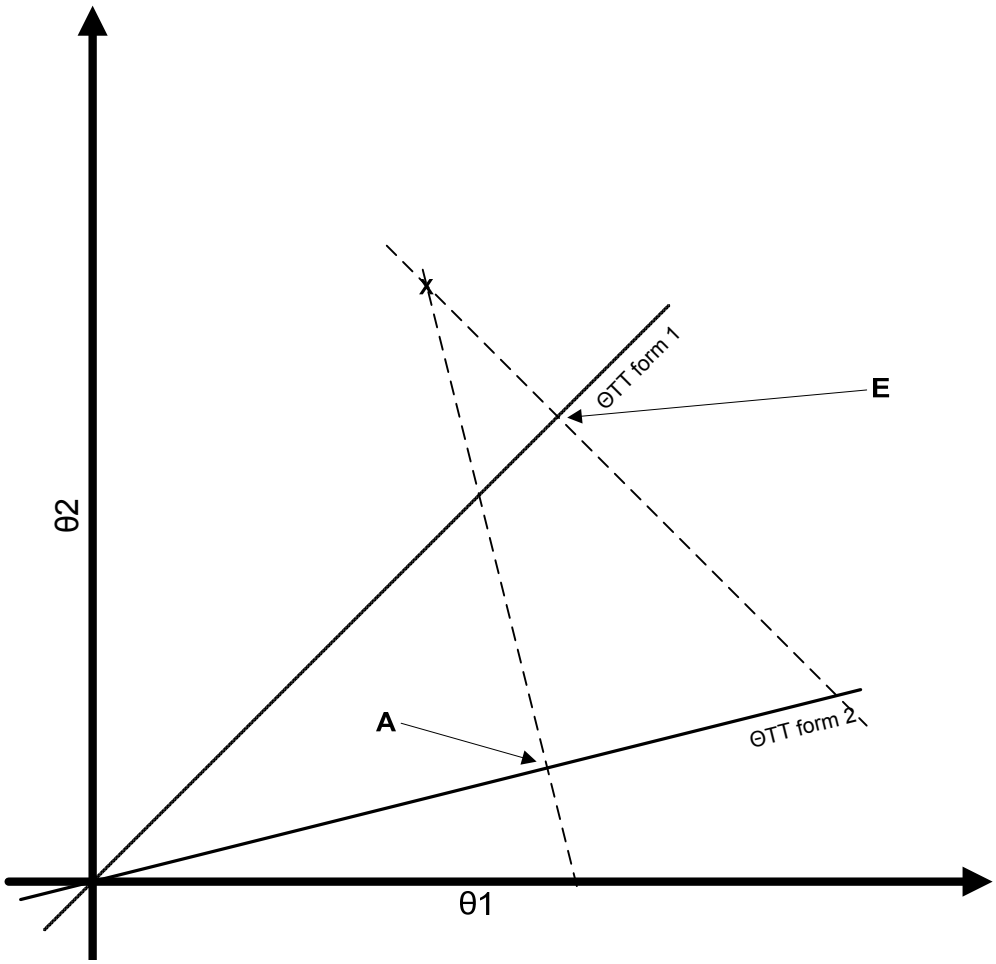

Figure 7c: Unidimensional equating in the multidimensional $\theta_1\theta_2$ plane

An examinee x with a two-dimensional ability (θ_1,θ_2) is associated with a certain unidimensional ability on form 1, labeled E in Figure 7c (consistent with Figure 6), and a certain (different) unidimensional ability on form 2, labeled A. The value of E and A depends on (θ_1,θ_2) as well as the number of items measuring each of the abilities on both forms – it's the average of (θ_1,θ_2) weighted by the number of items on each form.

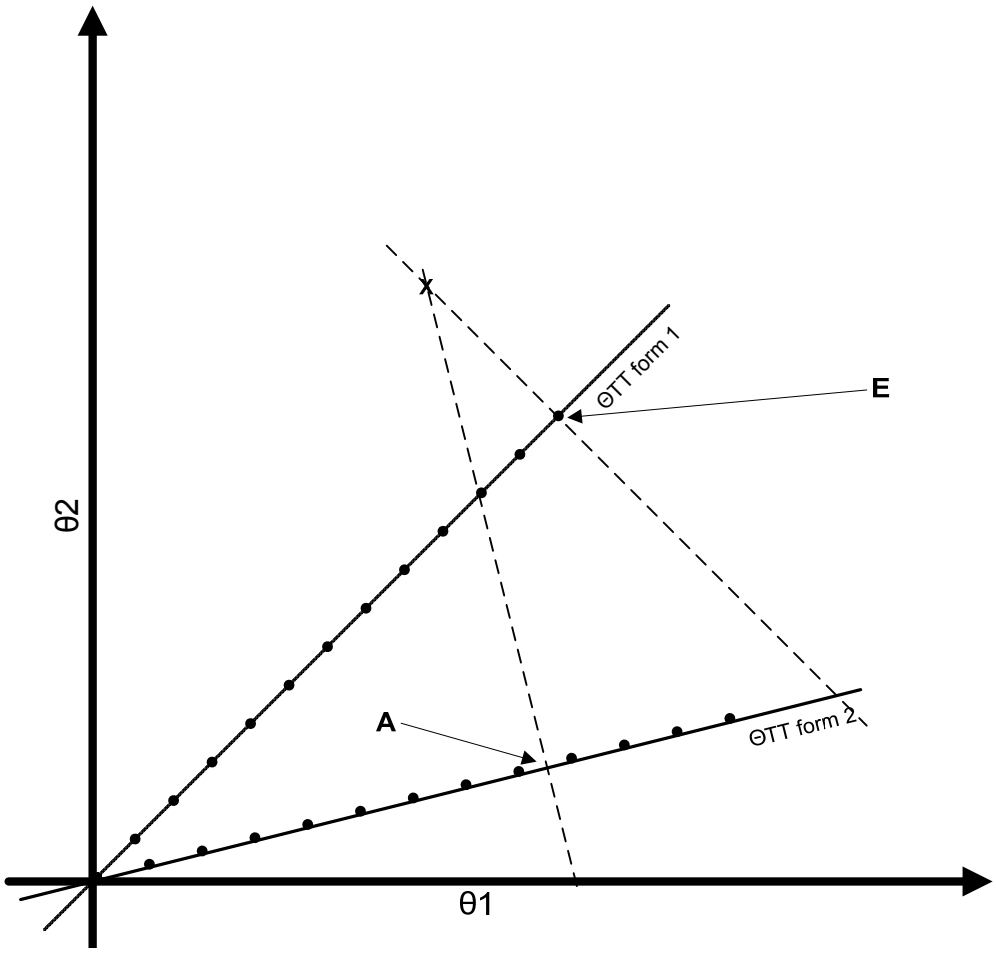

Figure 7d: Unidimensional equating in the multidimensional $\theta_1\theta_2$ plane

The length of the dotted line (12 dots from the origin) in Figure 7c indicates the Θ_{TT} of examinee x on form 1 (the original form). The same length is used along the $\Theta_{TTform\ 2}$ (the form being equated), since it is assumed the examinee will have the same unidimensional true ability, regardless of the form she takes. For this particular examinee (who's magnitude of θ_2 is slightly higher than of θ_1 as determined by the position of the x) this is longer than this examinee's unidimensional ability form 2 (which, as stated above, is A).

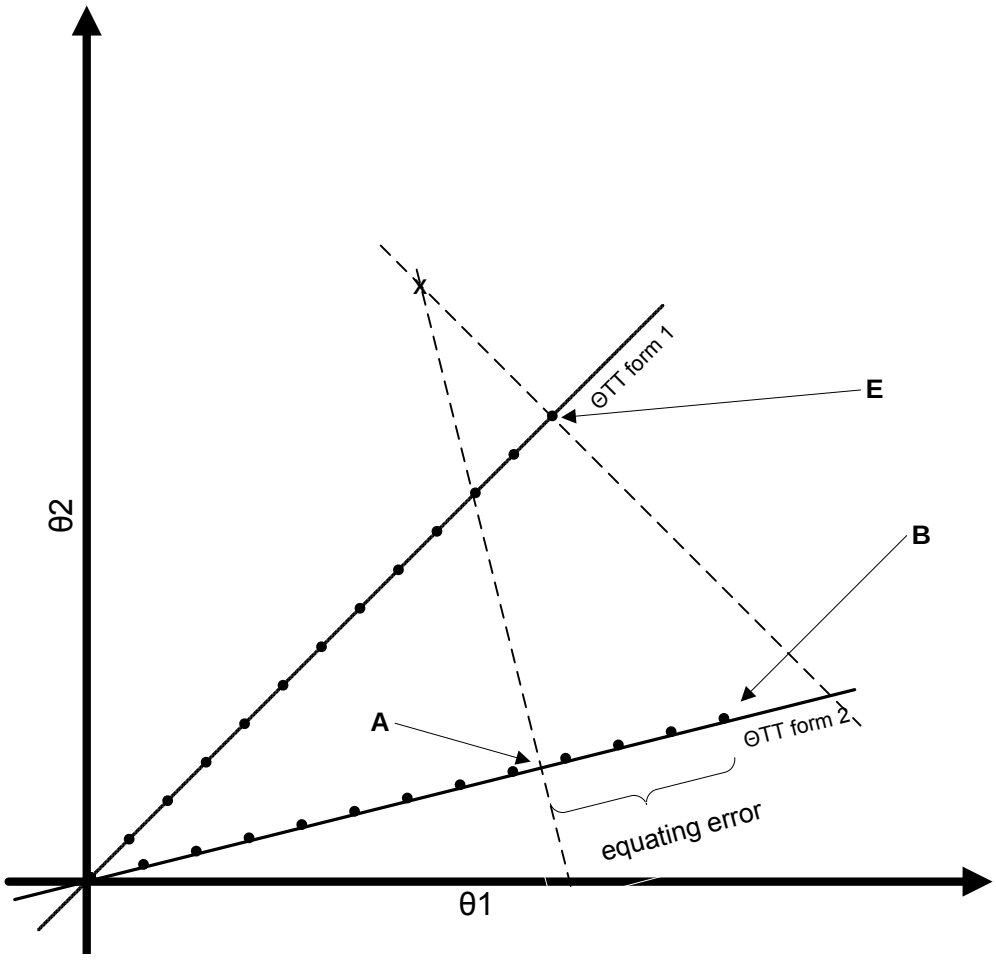

Figure 7e: Unidimensional equating in the multidimensional $\theta_1\theta_2$ plane

The difference between the unidimensional ability reflected on form 1 vs. the "true" unidimensional ability as calculated on form 2 is the error of equating form 2 to form 1. In other words - what true score did an examinee with a given multidimensional (θ_1,θ_2) get equated to vs. what score should they have been given. In this particular example, if form 2 were to be equated to form 1, x would have gotten an equated true score of B

($eq\tau_{f2}$) instead of a lower score of A ("true" true score, $t\tau_{f2}$) because of the change in item mixture of the forms.

3.3 Evaluating results

Both forms were calibrated (separately) using Parscale (code is included in Appendix D). This resulted in each simulating examinee obtaining two abilities for each of the forms – the calculated ability and the ability estimated from Parscale. The differences between the "true" unidimensional ability (T_2, point A in Figure 7e) and equated true score (eq_2T_2, point B in figure 7e) for the calculated and estimated abilities were calculated using the indices described in chapter 2.3. Those indices are naturally weighted since they are summed across all examinees.

Depending on which ability was used in calculation of the indices, Mean Signed Difference and Root Mean Squared Difference were calculated using abilities calculated as described in the previous section and abilities obtained from Parscale estimation (calculated MSD and RMSD and estimated MSD and RMSD (called MSDE and RMSDE respectively)). For each setting, the overall error was calculated as well as the errors for unidimensional top and bottom 10 percent of the population. For the settings in which half of the population differed on the second dimension, the top and bottom 10% of each of the abilities was considered as well as the errors for each of the population halves. The errors for those were also calculated for the no difference in subpopulation means setting; however, those are not discussed in detail here under the assumption that errors are symmetric in this condition.

As a crude estimate of the top and bottom distribution errors, the unidimensional 10% of the population were calculated. Those cutoffs were taken from the univariate

normal distribution – for the no difference between subpopulations settings, if θ_1 was lower than -1.28115 or higher than 1.28115 it was summed as bottom or top 10% of θ_1 respectively. A similar calculation was done for θ_2. If both of the abilities were lower or higher than those cutoffs the top/bottom 10% combined was calculated, if only one of the abilities was above/below the cutoff, the errors were classified at that ability. For the 0.5 difference the respective cutoffs were -1.03155 and 1.53155 (corresponding to the univariate normal distribution with mean of 0.25); for the 1 difference between subpopulations the cutoffs were -0.78155 and 1.78155 (corresponding to the univariate normal distribution with mean of 0.5).

Note that the 20-20:20-20 setting was possible for estimated abilities only, since the weighted ability for form 1 is the same as the weighted calculated ability for form 2, all errors for the calculated abilities in this setting would have been zero; consequently, they are not used in any analyses.

In order to examine which of the factors of interest (correlation between traits, form divergence and difference in subpopulation means) were the most important in determining the magnitude of the equating errors, a MANOVA analysis was run in SAS to model MSD, MSDE, RMSD and RMSDE as a function of the factors above, which were treated as fixed effects. The analyses were split for results for the full population and top and bottom 10% of the population. The mean errors across all simulations (rather than replications) were used as dependent variables.

The design of the study (4 way factorial, with 3 primary factors of interest) directly answered the first research question about what influences the magnitude of equating errors for non-parallel test forms. The second research question was answered

by observing the magnitude of the equating errors for various indices and comparing them to the least divergent form or the baseline setting. Answers to the first two research questions informed the answer to the third research question.

3.4 Algorithm for the study

For a quick reference, the following algorithm for the study is provided.

For each setting (model, correlation between traits, divergence of forms, divergence of population means) and every simulation do the following:

1. simulate data – simple structure two two-dimensional forms. Each form is simulated separately, but the same vector of examinee abilities is used to obtain the responses.
2. for each of the forms calculate the reference composite ("true" unidimensional ability) on each of the forms as the weighted average of the two unidimensional abilities, where weights are defined as the number of items measuring each ability for each subscale.
3. run both forms through Parscale (independently) – estimate the examinee unidimensional ability and item parameters, on both forms.
4. calculate "true" true score on form 2 – this is the sum of probabilities of responding to the item using the ability on form 1 (calculated in 2) and estimated item parameters on form 2, depicted by point As in Figure 7e.

$$T_2 = \sum P(u = 1 \mid \hat{F}_2, \Theta_1)$$

Θ_1 scaled to (0,1), $\hat{F}_2$ are the estimated item parameters on form 2.

5. calculate the equated true score on form 2 – this is the sum of probabilities of responding to the item using ability on form 2 (calculated in 2) and estimated item parameters on form 2, depicted with point B in Figure 7e.

$$\hat{T}_2 = eq_2(T_2) = \sum P(u=1 \mid \hat{F}_2, \Theta_2)$$

Θ_2 scaled to (0,1)

6. the difference between 5 and 4 above is the equating error.

$$\varepsilon = \hat{T}_2 - T_2$$

7. for Parscale estimated abilities, repeat steps 4,5 and 6, but use estimated unidimensional ability on forms 1 form 2, respectively.

$$t_2 = \sum P(u=1 \mid \hat{F}_2, \hat{\theta}_1)$$

Estimated θ_1 scaled to (0,1)

$$\hat{t}_2 = eq_2(t_2) = \sum P(u=1 \mid \hat{F}_2, \hat{\theta}_2)$$

Estimated θ_2 scaled to (0,1)

$$\hat{\varepsilon} = \hat{t}_2 - t_2$$

8. calculate error indices for the full population:
    a. mean signed difference – sum the differences calculated in 6 for all examinees, divide by the number of examinees.
    b. root mean squared difference – square the absolute value of the differences in 6, then sum across all examinees, divide by the number of examinees and take the square root.
9. calculate error indices for the parts of the population of interest – repeat 8a and 8b, but calculate the sums only for the following examinees:
    a. theta 1 and theta 2 ability each above 1.281552 (1.531552 and 1.781552 for higher differences in the means)
    b. only one of the abilities above the values in a)
    c. theta 1 and theta 2 ability each below -1.281552 (-1.031552 and -0.7815516 for higher differences in the means)
    d. only one of the abilities below the values in d)
    e. subpopulation 1 examinees
    f. subpopulation 2 examinees

    keep the counts of how many examinees are in each of the groups, divide by the number of appropriate examinees belonging to each of the groups.
10. save 8 and 9 into a vector, calculate the mean across all simulations.

4 Results

The results are structured as follows - first, a general description of the 3PL model results is provided for no difference in the mean and for the differences in the mean (0.5 and 1) conditions. The behavior of the indices is described for the full population, top/bottom 10% of the full population, top/bottom of each of the abilities and for each of the subpopulations (subpopulation 2 is the one with increased means on the first dimension). Values of the errors are provided for some settings for illustration purposes; the full result tables are included in Appendix A. Last, the results of the ANOVA analyses for MSD, MSDE, RMSD and RMSDE are presented with a discussion of the results. The behavior of the Rasch model is also included in those results. The full SAS output of proc glm is included in Appendix B.

Because the equated true score was subtracted from the true score based on ability estimates, negative MSDs and MSDEs indicate that the equated true score was larger than the true score based on the calculated and estimated ability – i.e. the equating overestimates the true score and gives higher score to the examinees than they would have gotten if they had taken a different form.

Generally, the intuition holds – the scoring errors increase with the decreased correlation between traits and increased disparity between forms. However, there are some exceptions to this and some trends which are interesting to note which are discussed below.

True score equating overestimates at the bottom of the combined ability scale and underestimates at the top regardless of the subpopulation mean difference. This under- and over-estimation respectively seems to be due mainly to the errors in equating of the

θ_2 ability where the errors are biggest. For the bottom 10 percents of each of the abilities, the errors are always higher for θ_2 than for θ_1 (in absolute magnitude for the MSD and MSDE). Errors for the subpopulation with all zero ability are very small and positive for all settings; in contrast, errors for the population half with the mean ability of 0.5 and 1 were much larger and positive – it seems that the more able half of the population would have actually been penalized with equating. The effect of the differences in the population means on the equating errors is much larger when 0.5 difference is compared to difference of 1 than when 0.5 difference is compared to the difference of 0.

As more examinees become available at higher abilities with the shift in means, the errors decrease for those groups. The opposite is the case for the bottom of the ability ranges.

The correlation of 0.9 seems to exhibit the most behavior which is contrary to intuitive trend. Removing 5 items also shows that kind of behavior for many settings. This could be explained by the high dependence of the errors on the quality of items removed for 1 or 5 items, i.e. removing one strong item (highly discriminating with most information for the population) will have more effect on the true score than removing 5 not as informative items. Errors for the most divergent forms are also often smaller then the errors for the less divergent forms. This is especially pronounced for the extremes of the population.

Based on the indices using estimated abilities, one could say for instance that removing 5 items at the 0.5 correlation with no difference in subpopulation means has the same effect on the equating errors as completely switching subscales for the 0.95 correlation between traits. However, this result is not confirmed for the calculated

abilities. At correlation of 0.5 removing all 20 items from the subscale has the same effect in terms of mean absolute difference error as switching subscales at correlation of 0.7.

Most of the variability in errors is due to forms diverging rather than the change in correlation. This is especially true for the non zero differences in subpopulation means setting.

Generally, the errors in the Rasch model are larger than for the 3PL model. In particular, the Rasch model does not perform well with fewer examinees to estimate – for the bottom 10% of each of the abilities with non-zero mean, the errors are larger than for the 3PL. This is probably due to the guessing parameter of the 3PL model which reduces the possible variation in the score on either form. However, for some settings the errors in Rasch equating are smaller; the equating does not overestimate as much as the 3PL does. As the subpopulations diverge, the errors become similar to the 3PL for all the population parts examined but the bottom of each ability (where the sparseness of the data seems to effect the Rasch model more than the 3PL model).

The RMSD and RMSDE errors differ rather substantially in magnitude, especially for the low divergence between the forms. The exact reason for this discrepancy is unknown at this time; however, closer investigation of the linearity of the calculated approximation of the reference composite could shed some light on this issue.

For convenience, the following convention was adopted for naming the different form settings: a-b:c-d where a is the number of items on the first form, first dimension, b is the number of items on the first form second dimension, c is the number of items on

the second form, first dimension and d is the number of items on the second form, second dimension.

The calculated setting do not allow for the 20-20:20:20 error comparison – the difference between eqτ_2 and τ_2 is zero for this setting since the weighted reference composite of both the first and the second form is the same.

4.1 Main settings results – no difference in population means

For this setting, by design, the top 10% of both abilities is symmetric with the bottom 10% of both abilities – for MSD and MSDE they differ in the sign, for RMSD and RMSD they are similar in magnitude. The same is true for the top and bottom 10 percent of each of the abilities. The errors for each of the subpopulations are the same as for the full population. Those results are included in the Appendix A, but are not discussed here.

MSD. The MSD errors for the full population are very small (zero to one decimal place). The top 10% combined errors are mostly negative and increase as forms diverge for every correlation till the 0-40:40-0 setting when they are almost as small as for the 20-20:21-19 setting. As the correlations increase, the errors decrease. Most of the top 10 percent of θ_1 indices are positive (with very slight overestimations for 20-20:40-0 setting); the general trend holds – indices increase as forms diverge and decrease as correlations increase.

MSDE. The mean signed difference for the whole population for estimated abilities is relatively small and negative. As forms diverge, the errors increase (with the exception of 20-20:40-0 vs. 0-40:40-0 setting, where they decrease slightly). Overall, as

the correlations increase the indices decrease, but there seems to be little difference between the 0.7 and 0.9 correlation. With the shift to 0.95 from 0.9 correlation the indices decease rapidly. The general trends are the same as for the MSD.

RMSD. The trend is kept without exceptions for the full population – the RMSD ranges from the high of 7.17 for 0-40:40-0, 0.5 correlation setting to 0.067 for the 20-20:21-19, 0.95 correlation.

RMSDE. The trend is kept, although it seems that most of the trend can be explained to the divergent forms rather than the decrease in error as the correlations increase (this is confirmed with the glm analysis in section 4.4). For example, for the full population, for the baseline setting, the errors decrease from 3.34 to 3.11 as correlations increase from 0.5 to 0.95, but for 0.5 correlation, the error increases more than two-fold from 3.34 to 7.52 between the baseline and the 0-40:40-0 setting. For the higher correlations (0.9 and 0.95) this increase is not as pronounced (from 3.36 to 4.38 and from 3.11 to 3.78 respectively). The errors for the top 10% of the population are inconsistent with higher divergence between the forms, sometimes exhibiting higher errors than the baseline setting. For the top/bottom of each of the abilities, the errors are much larger than for the top of the overall population (e.g. 2.25 vs. 4.27 for the 0.7 correlation 20:20’40-0 setting).

4.2 Different population means

For this setting, half of the population (subpopulation 2) mean was set to either 0.5 or 1 for the first dimension (θ_1). The standard deviations of the abilities were kept at 1 for either half of the population. The populations were combined for Parscale estimation.

MSD. All the mean signed differences for the full population are positive. As the forms diverge, the equating errors for the full population increase for both differences in the means. As correlations increase however, for the 0.5 difference between means, the errors seem to stay relatively constant. For the 1 difference between means the errors increase with correlation increase. For both abilities in the top 10% range, the errors are very small for one and five items removed for both differences in population means. For 20 items removed all errors are negative decreasing from -0.843 with 0.5 correlation to -0.421 with 0.95 correlation for 0.5 difference between means and -0.44 to -0.17 for 1 difference between means. For both abilities in the bottom 10% range the errors are positive and relatively constant across correlations for 0.5 difference in means, the trend is kept for 1 difference in means. The errors for the 0-40:40-0 setting are lower than for the 20 items removed, increasing from 0.253 to 0.646 for the minimum and maximum of the examined correlations for 0.5 difference between subpopulation means, and from 0.349 to 0.804 for 1 difference in subpopulation means. Equating overestimates the true score for top 10% of θ_2 with contrary to trend errors for the most divergent forms and higher correlations. In addition, for the top 10% of θ_2 for the two most divergent forms, all of the differences are negative and smaller (in absolute magnitude) than the top 10% of θ_1 for both differences in means. For the bottom 10% of θ_1 errors are negative – they become more negative as forms diverge and closer to zero as the correlations increase for 0.5 difference in means. For 1 difference in means, errors are larger than for the no difference and 0.5 difference for all settings and trend is kept. For the bottom 10% of θ_2 all mean signed differences are positive and much larger (in absolute terms) than the top 10% θ_2. The absolute magnitude to MSD is larger for the bottom 10% of both abilities

than for the top 10% of both abilities. Errors for subpopulation 1 are very close to zero, while the errors for subpopulation 2 are much larger (3.517 for 0.5 correlation most divergent forms, 0.5 difference, 7.05 for 1 difference). The trend is kept with errors increasing as forms diverge for subpopulation 2 and decreasing as correlations increase, however, this decrease is not as pronounced across correlations.

MSDE. The overall errors for the full population are relatively small (from slightly negative -0.123 for one item removed 0.5 correlation between traits to 0.36 for the 0-40:40-0 0.5 correlation setting for 0.5 difference in means). Generally they increase as forms diverge and decrease as correlations increase with errors for the last setting smaller for low correlations (0.5 and 0.7) between traits than errors for the 20-20:40-0 setting. Trend is kept for 1 difference in means for all settings but the most divergent one. For the top 10% of both abilities for both differences in means the errors are again small and decreasing rapidly as correlations increase, but decreasing as forms diverge for 0.5 correlation. The bottom 10% of both abilities the errors increase as forms diverge and decrease as correlations increase. For the most divergent forms the errors decrease when compared to the 20-20:40-0 setting. Again, as with MSD, most errors are positive for top 10% of θ_1 and negative for top 10% of θ_2; the reverse is true for bottom 10% of each ability. For 1 difference in the means the absolute magnitudes are similar with no apparent trend for lower correlations, while for the higher correlations the top 10% θ_2 is slightly higher than the θ_1. For the bottom 10% of θ_1 the errors are all negative; they are all positive (and bigger in absolute terms) for bottom 10% of θ_2. As before, equating underestimates the true score for the top abilities of θ_1 and overestimates for bottom abilities; the reverse is true for θ_2. Same trends as with the MSD can also be seen for the

subpopulations with subpopulation 2 exhibiting larger (and all positive) errors than subpopulation 1.

RMSD. The trend is kept for the full population and for the top 10% of both abilities. For the bottom 10% the errors are generally larger than the ones for the top 10% of the overall population, much larger for the larger difference in the subpopulation means. For both differences in the means the trend is kept for top and bottom 10% of each of the abilities. The errors for top 10% of θ_1 are higher for the 20-20:40-0 setting than the top 10% of θ_2 for both differences in population means. They are comparable for 1 and 5 items removed. The opposite is true for bottom 10% of θ_1 – here the errors are much smaller than those for bottom 10% of θ_2. The magnitude of the errors is similar for the top and bottom 10% of θ_1 for 0.5 difference in subpopulation means. They are slightly higher for 1 difference in the means. For θ_2, bottom 10% is much higher than the top 10% for both differences in the means. The trend is kept when both population halves are examined separately; however the errors are higher for subpopulation 2 than for subpopulation 1. The differences are more pronounced the more divergent the forms and seem more pronounced with lower correlations (7.41 vs. 8.07 for 0.5 correlation and 2.48 vs. 4.57 for 0.95 correlation for 0.5 difference in the means).

RMSDE. The trend is kept for the full population. For the top 10% and the bottom 10% of both abilities, the errors decline slightly for the five items removed setting. For larger difference in subpopulation means, the baseline errors are higher than one and five items removed. For the bottom 10% the errors increase slightly as correlation increases from 0.7 to 0.9 for the baseline and moderately divergent forms to drop again with 0.95 correlation between traits for 0.5 difference in the means. For large difference in

subpopulation means the errors are slightly smaller for top 10% of θ_1 than for the top 10% of θ_2. The same is true for bottom 10% for each of the abilities individually (everything but the baseline at 0.7 correlation). Errors are larger for the bottom 10% of both abilities than the top 10%, bottom 10% of θ_1 than top 10% of θ_1 and much larger for bottom 10% of θ_2 (for the most extreme settings in terms of form divergence) than for the top 10% of θ_2. The two subpopulation errors reflect the general trend, but there are no noticeable differences in the magnitude of the errors for 0.5 difference in the means; for large difference in subpopulation means the errors are slightly smaller for subpopulation 2 at low form divergence, but larger at higher form divergence.

4.3 Rasch model results

In this section comparisons are made between the errors of the 3PL model and the errors in the 1PL model. Where examples are given, the errors for the 1PL model are listed first.

No difference in population means

Overall results. The MSD and MSDE errors are slightly larger, in absolute magnitude than the corresponding 3PL errors. RMSD and RMSDE errors are larger as well.

Top 10% combined. The errors are larger here than for the 3PL, there is also a drop at the most divergent setting, like with the 3PL, the errors are negative. MSDE directionality is even less consistent than for the 3PL model, no conclusions can be drawn about the magnitude of the errors (they are generally rather small). RMSD errors are very close in magnitude to the 3PL; RMSDE on the other hand are generally smaller.

Top 10% θ_1. Trend is kept, most MSD are positive and similar in magnitude as for the 3PL for one and five items removed, but slightly smaller for the more divergent forms. MSDE is mostly positive and smaller than for the 3PL (e.g. 0.736 vs. 1.665 for 0.5 correlation five items removed, 0.946 vs. 1.024 for 0.95 subscale removed) as are the RMSDE. RMSD magnitudes are relatively comparable between the models.

Different population means

Overall results. The MSD 1PL errors are slightly larger than the 3PL errors, but the difference is minimal (for instance 1.81 vs. 1.82 for 0.7 correlation, 0-40:40-0 setting for 0.5 difference in means, 1.24 vs. 1.41 for 0.5 correlation one subscale removed for 1 difference in means). The MSDE is smaller for most settings particularly at lower correlations for both differences in means. RMSD and RMSDE are also slightly larger than for the 3PL.

Top 10% of both abilities. For both differences in means, MSDs are comparable in magnitude to the 3PL for the small divergence and slightly larger (in absolute magnitude) for the higher form divergence. For MSDE most errors are negative. For 0.5 difference in means approximately as many errors are larger as are smaller than the 3PL in terms of absolute values, for 1 difference in means the absolute value of the errors is larger than for the 3PL model. RMSD and RMSDE are not noticeably different for the 1PL than for the 3PL for 0.5 difference in means, they are slightly smaller for the Rasch model than for the 3PL for 1 difference in means.

Bottom 10% of both abilities. MSD errors are much bigger than for the 3PL model. Again, the errors remain more or less constant across the correlations. For MSDE,

the absolute magnitude of the errors is greater than for the 3PL except for the baseline condition 0.5 and 0.95 correlation and the most extreme condition 0.7 and 0.95 correlation for 0.5 difference in means. For 1 difference in means MSDE is larger than for the 3PL errors and mostly positive. RMSD errors are much larger for the 1PL than for the 3PL (e.g. 0.5 correlation most divergent forms 4.54 vs. 2.68 for RMSD for 0.5 difference in means. The respective number for 1 difference in means is 5.6 vs. 3.5). RMSDE are also larger.

Top 10% of θ_1. MSDs are smaller for more divergent forms and the same for less divergent form for 0.5 difference in the means. Reverse is the case for 1 difference in the means. MSDEs are mostly positive and smaller than those of the 3PL for all but the highest correlations least diverse settings (baseline and 1 item removed) For 1 difference in means the only exceptions are one item removed correlation 0.7 – 0.257 vs. 0.153 and baseline correlation 0.95 – 0.073 vs. -0.043. RMSD errors are slightly smaller than for the 3PL. Same is true for the estimated root mean square differences for 0.5 difference in means. For 1 difference in means RMSDE is generally larger than for the 3PL, however those differences are not big. The error for 0.7 correlation one subscale removed being exceptionally small (smaller than for the five items removed for this correlation).

Top 10% of θ_2. The MSD errors are negative and mostly larger (i.e. more negative) in absolute magnitude for both differences between the means. MSDE are mostly negative and slightly smaller (i.e. larger in absolute magnitude) than the 3PL for 0.5 difference in subpopulation means. They are larger (more negative) for more divergent forms and smaller for less divergent forms and the baseline for 1 difference in subpopulation means. RMSDs are comparable to the 3PL for low form divergence and

smaller for high form divergence for both differences in the means. RMSDE errors are slightly smaller than for the 3PL for 0.5 difference in population means, for 1 difference RMSDE is smaller for baseline and one and five items removed, but bigger for most divergent forms.

Bottom 10% of θ_1. Both MSD and MSDE errors are larger (more negative) than for the 3PL. RMSD and RMSDE are larger than for the 3PL for both differences in subpopulation means.

Bottom 10% of θ_2. For both subpopulation differences, the MSDs are larger than for the 3PL model. MSDEs are larger for higher form divergence and slightly smaller for low form divergence and the baseline setting. RMSDs are larger than for the 3PL (e.g. 0.7 correlation one subscale removed 7.54 vs. 6.42 for 0.5 difference in the means, 8.25 vs. 7.5 for 1 difference in the means. RMSDE is also larger.

Subpopulation 1. MSD is generally very small. The 20-20:40-0 error is much larger (decreasing from 0.126 to 0.076 across correlations for 1PL and increasing from 0.118 to 0.215 for the 3PL for 0.5 difference in subpopulations and increasing from 0.241 to 0.28 for 1PL vs. 0.017 to 0.18 for the 3PL model for 1 difference in subpopulations). For MSDE for the two most divergent settings the errors are bigger (more negative) for most correlations (with the exception of the 0.9 correlation where the errors are smaller for 0.5 difference in subpopulation means and 0.95 correlation baseline, five and 20 items removed and 0.9 correlation baseline for 1 difference in subpopulation means); for the least divergent settings (including the baseline) the errors are smaller (with the same exception of 0.9 correlation). RMSD and RMSDE are also larger for the 1PL.

Subpopulation 2. For 0.5 difference in means MSD is larger for subpopulation 2 for 1PL than for 3PL especially for more divergent forms. MSDE is slightly smaller in magnitude for both subpopulation mean differences. The differences are slightly more pronounced for low form divergence with 1 difference between means. RMSD and RMSDE are larger than for the 3PL.

4.4 GLM results

As can be seen from sections 4.1 though 4.3 and in Appendix A, the sheer number of results is overwhelming. In order to better understand the factors at work, a generalized linear model was run on the errors. This chapter presents the results of this model. Form difference, correlation and difference in population means were treated as fixed factors and used in full interactions. Whether the error came from calculated or estimated reference composite was also included in the model, but not used in interactions.

The SAS' proc GLM output full population results for MSD (bias) and RMSD are included in Appendix B. Table 3 below lists the factors indicated by the model as significant at 5% level.

Table 3: Factors significant for the magnitude of equating errors.

Portion of population	3PL		1PL	
	MSD	RMSD	MSD	RMSD[#]
Overall	Calculation vs. estimation Form difference Difference in pop. means Form difference*Pop.difference	Calculation vs. estimation Form difference Correlation Correlation*form difference	Calculation vs. estimation Form difference Difference in pop. means Form difference*Pop.difference	Calculation vs. estimation Form difference Difference in pop. means Correlation
Top 10% of both abilities	Form difference Difference in pop. means	Calculation vs. estimation Form difference Difference in pop. means Correlation	Form difference	Calculation vs. estimation Form difference
Bottom 10% of both abilities	Calculation vs. estimation Form difference Difference in pop. means	Calculation vs. estimation Form difference Difference in pop. means	Calculation vs. estimation Form difference Difference in pop. means	Calculation vs. estimation Form difference Difference in pop. means

- Generalized Linear Model is not significant at 5% level for the top 10% of both abilities

The r-squared for the models are given in table 4 below:

Table 4: R-squared for the respective models

Portion of the population	3PL		1PL	
	MSD	RMSD	MSD	RMSD
Overall	0.73	0.92	0.73	0.85
Top 10%	0.81	0.92	0.72	0.55
Bottom 10%	0.85	0.86	0.86	0.83

Large proportion of variability in the data is explained by the model, as depicted in Table 4. This indicates that the errors in equating as measured by this study can in fact be predicted fairly well by the factors examined. From Table 3 one can see that the form difference and difference in population means are the most decisive factors in predicting the magnitude of the errors. Somewhat surprisingly, correlation is important only for the RMSD overall results. In addition, whether the reference composite used was calculated or estimated is also significant. Very similar results (in terms of significance of the remaining factors) were obtained when the variable indicating calculated or estimated error was not included in the model.

An illustration of the form divergence significance on the overall results as measured by the bias is in Tables 5 and 6 below (identical to Tables A1 and A19 in the Appendix).

Table 5: MSD using calculated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.000	-0.001	0.000
25-15	-0.004	0.012	0.001	0.003
40-0	-0.045	0.008	-0.005	0.033
0-40:40-0	0.038	-0.037	-0.007	-0.001

Table 6: MSD using calculated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.047	0.049	0.048	0.048
25-15	0.215	0.235	0.236	0.234
40-0	0.760	0.806	0.820	0.855
0-40:40-0	1.746	1.814	1.799	1.829

In Table 6, the bias (MSD) increases as the forms diverge (from 0.05 to 1.745 for 0.5 correlation and 0.05 to 1.8 for 0.9 correlation), but stays more or less the same as correlations increase for the same divergence between the forms. In Table 5, the errors are much smaller than in Table 6, illustrating the significance of the difference in population means. In Table A37 in the appendix (overall results, difference of 1 between subpopulation means) the errors are even bigger than the errors for 0.5 difference above. To illustrate the importance of the interaction factor, one has to look across Tables 5 and 6. One can see that the increases in errors are not the same for different population

means. For example, errors increase tenfold from 20-20:25-15 setting to 20-20:40-0 setting for no difference between means, but only 3.5 times for 0.5 difference between means. Similar results can be seen in Tables A37 (for 1 difference in the means), A109, A127, A145 for the estimated reference composite, and Tables A55, A73, A91, A163, A181 and A195 for the 1PL model).

In a manner similar to the above, one can illustrate the RMSD differences with Tables 7 and 8 below (identical to Tables A28 and A136 in the Appendix). The difference in the errors, and thus the significance of the factor, between the calculated (Table 7) and estimated (Table 8) reference composite is clearly visible. The errors increase with decreased correlation and increased divergence between forms. For the interaction term, one can see in Table 8 that the ratio of the 20-20:21-19 form divergence 0.5 to 0.95 correlation is 1.14 (3.464/3.027=1.14), while the ratio for the most divergent form is 1.85. Notice however, that for the errors calculated using the calculated composite, those ratios remain very close, ranging from 2.12 for the 20-20:40-0 setting to 2.28 for the 20-20:21-19 setting.. This might be an indication of a 3-way interaction between the variables.

Table 7: RMSD using calculated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.224	0.175	0.116	0.098
25-15	1.053	0.852	0.590	0.479
40-0	4.138	3.420	2.252	1.943
0-40:40-0	7.754	6.462	4.346	3.640

Table 8: RMSD using estimated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.323	3.011	3.071	3.019
21-19	3.464	3.062	3.023	3.027
25-15	3.501	3.183	3.116	3.106
40-0	4.954	4.285	3.571	3.375
0-40:40-0	7.753	6.548	4.776	4.172

The RMSD analysis for errors for overall results for the 3PL model can be seen in Tables A10, A46 (for the calculated latent composite) and Tables A119 and A154 in addition to the ones mentioned above.

The next set of tables refers to the top 10% of the overall population MSD error predictors.

Table 9: MSD using estimated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.261	-0.032	0.447	-0.006
21-19	-0.068	-0.053	0.020	0.124
25-15	0.203	-0.052	-0.051	-0.006
40-0	-0.751	-0.380	-0.230	-0.098
0-40:40-0	0.119	-0.049	0.024	0.006

Table 10: MSD using estimated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.343	-0.078	0.010	0.001
21-19	0.213	-0.007	-0.013	-0.019
25-15	0.199	-0.011	0.014	0.003
40-0	-0.472	-0.406	-0.236	-0.128
0-40:40-0	-0.004	0.199	0.025	0.029

For the top 10% of the population for the 3PL model compare Table 9 (Table A110) and Table 10 (Table A128) above – with few exceptions, as the forms diverge, the errors increase. As the population means diverge, the errors also increase for most of the settings. The decrease in errors as correlations increase is not significant, probably because of certain inconsistencies of trend, especially visible in Table 10. Errors in this table actually increase with some increased correlations.

Because some errors increase and others decrease for the same form divergence across different population means, it's easy to see why the interaction term was not significant in the glm analysis. For instance the ratios of the 0.5 correlation error to 0.95 correlation error for a 20-20:21:19 form distribution is 0.51 (i.e. and increase in error) for no difference in population means and 11.21 for 0.5 difference in means (i.e. a decrease in error).

The MSD and MSDE errors illustrating the same points for the 3PL model can be seen in Tables A2, A20, A38 (for the calculated reference composite) and Table A146, in addition to the tables above (for the estimated reference composite).

For the illustration of the importance of the correlation for the RMSD, please refer to Table 11 below (Table A11). A significant drop in error is evident as the correlations between abilities increase. As the forms diverge, the errors increase. Similar results for the 3PL model can be found in Tables A29 and A47 (for the calculated reference composite) and Tables A119, A137 and A155 (for the estimated reference composite)

Table 11: RMSD using calculated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.044	0.039	0.032	0.022
25-15	0.258	0.223	0.181	0.125
40-0	1.796	1.553	1.200	1.011
0-40:40-0	1.939	1.570	1.393	0.982

For the bottom 10% of both abilities, compare Table 12 (Table A23 in the Appendix) to Table 13 (Table A41 in the Appendix).

Table 12: MSD using calculated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.010	0.012	0.017	0.017
25-15	0.160	0.148	0.135	0.151
40-0	2.068	1.620	1.490	1.635
0-40:40-0	0.253	0.431	0.542	0.646

Table 13: MSD using calculated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.015	0.016	0.021	0.025
25-15	0.193	0.190	0.175	0.199
40-0	2.408	2.142	1.840	1.808
0-40:40-0	0.349	0.646	0.806	0.804

In both tables, the errors increase with increased form divergence is evident (with the exception of the most extreme setting where the errors are actually smaller than for the 20-20:40-0 setting). In Table 12, the errors are higher than in Table 11, illustrating the significance of the difference in population means factor.

The MSD and MSDE errors illustrating the same points can also be seen in Table A5 (for the calculated reference composite), A113, A131 and A149 (for the estimated reference composite) for the 3PL model. 1PL model errors where the same factors proved significant, Tables A59, A77 and A95 (for the calculated reference composite) and Tables A167, A185 and A203 (for the estimated reference composite) illustrate the points made above.

The RMSD significant factors are the same for the bottom 10% of the full population as the MSD significant factors. They are therefore not discussed here in detail. Please refer to Tables A14, A31, A50 and A59, A86, A105 for the calculated reference composite for the 3PL and 1PL model respectively. The estimated reference composite errors are in Tables A113, A140, A158 and A176, A194 and A212 for the 3PL and 1PL models respectively.

Differences between the 1PL and the 3PL model:

As can be seen from Table 3, the 3PL significant factors differ slightly from the 1PL significant factors for the RMSD for overall population (no interaction term and difference in population means significant) and the top 10% of the full population (in both MSD and RMSD). For the top 10% of the full population, the predictive model is not significant for the 1PL model. It is therefore hard to draw any valid conclusions for the differences in significant factors indicated.

Tables 14 (Table A64) and Table 15 (Table A82) below, together with Table 7 above, address the differences between the 3PL and the 1PL model RMSD errors for the full population.

Table 14: RMSD using calculated reference composite full population, no difference in subpopulation means, 1 PL model

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.218	0.171	0.114	0.069
25-15	1.061	0.836	0.496	0.368
40-0	4.133	3.324	1.999	1.653
0-40:40-0	7.727	6.135	3.616	2.499

Table 15: RMSD using calculated reference composite full population, 0.5 difference in subpopulation means, 1PL model

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.226	0.185	0.125	0.106
25-15	1.160	0.942	0.637	0.516
40-0	4.410	3.627	2.699	2.208
0-40:40-0	8.391	6.669	4.435	3.792

With the difference in population means, the errors are bigger, especially for the more divergent forms. However, there is no interaction between the correlation and form difference – compare increase in errors from five items removed to full subscale removed at 0.5 correlation – 0.25 ratio. This is the same ratio for the increase in errors for those form discrepancies for the remaining 3 correlations.

For the top 10% compare the Tables 16 and 17 (Tables A74 and A92 in the Appendix) below. While some of the errors increase with increased difference (for the 20-20:21-19 most correlations and 20-20:40-0 form difference), the remaining errors decrease. It is this inconsistency, not present in the 3PL model, which makes the difference in population means not a significant factor in predicting the magnitude of the errors.

Table 16: MSD using calculated reference composite top 10% of the full population, 0.5 difference in subpopulation means, 1PL model

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.004	0.007	0.011
25-15	-0.064	-0.039	-0.009	0.004
40-0	-0.960	-0.697	-0.415	-0.422
0-40:40-0	0.218	0.228	0.456	0.490

Table 17: MSD using calculated reference composite top 10% of the full population, 1 difference in subpopulation means, 1PL model

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.006	0.008	0.010	0.010
25-15	-0.020	-0.003	0.021	0.024
40-0	-0.374	-0.320	-0.163	-0.168
0-40:40-0	0.315	0.322	0.384	0.371

In order to further explore the effect of changing each of the setting, the LSMEANS procedure was conducted for the 4 variables of interest. The data design precluded using interactions with the procedure. The LSMEANS for the 3PL model overall results are presented in Table 18 below, the 1PL model results and top/bottom 10% of examinees results are in Appendix B with the rest of the SAS output.

Table 18: LSMEANS differences for 3PL model, overall results

Variable and levels		Statistic	
		msd	rmsd
	calculation	0.645	2.254
	estimation	0.159	3.898
correlation	0.5	0.430	3.810
	0.7	0.392	3.288
	0.9	0.397	2.707
	0.95	0.390	2.497
form difference	20-20	0.253	2.333
	21-19	0.023	1.639
	25-15	0.161	1.985
	40-0	0.544	3.540
	0-40:40-0	1.029	5.881
difference in pop. means	0	0.068	2.882
	0.5	0.414	3.018
	1	0.724	3.327

We can be seen in Table 18, the Mean Signed Difference decreases from 0.65 to 0.16, or four-fold when the abilities considered are estimated vs. calculated. For correlation, the errors decrease slightly with increased correlation between 0.5 and 0.7, then increase very slightly to decrease again. The cumulative effect of those decreases is not enough to make the result significant (see Table 3). The form difference shows big shifts in means – from a low of 0.023 with 1 item removed, to a high of 1.029 with a subscale shift. Generally, the errors increase with increased form discrepancy, it's interesting to note that the errors increase 1.9 times from 20 items removed to 40 items removed is. MSD increases by 0.32 for every 0.5 shift in population means. For the RMSD the decreases in errors due to increased correlation are more pronounced – the

error at 0.95 correlation is 65% that at 0.5 correlation. Comparing that result to the 1PL result (Appendix B), one can see that that rate of RMSD decrease is consistent for both models. A pattern similar to the MSD is also present for the RMSD – a drop for the 1 and 5 items removed as compared to the baseline setting and final increase, with slightly less than 1.9 times, in fact it's 1.7 times from 20 items removed to 40 items removed.

Similar comparisons can be conducted for other population parts and the 1PL model. However it's noteworthy that it would be impossible to tell if the effect is significant or not – note for instance the RMSDs for the difference in population means – the errors do increase, and yet the effect is not significant for this error. A different design of the study with variables increasing linearly would allow for setting up contrasts. Alternatively, if there were only two settings for each variable, an effect size could be calculated.

In conclusion – while there are some differences between the predictive factors of the magnitude between the 3PL and the 1PL errors those differences are not systematic in nature. Generally, the same factors are important for both models.

5 Final conclusions and further studies

5.1 Answers to the research questions

The following research questions were stated at the beginning of this study:

1. What influences the magnitude of the equating errors for multidimensional test forms:
    a. Correlation between dimensions – it was hypothesized that the higher correlation between the subscales, the more difficult it was going to be to distinguish between the dimensions on the test, which will result in smaller equating errors.
    b. Number of items on each subscale relative to the total number of items on the test – naturally, the more divergent the two forms of the test, the less equatable the tests.
    c. Different ability distributions of subpopulations on the dimensions. It was hypothesized that equating is more likely to fail with increased interaction of dimension and ability.
2. Conversely – when are the equating errors small enough to justify the same construct assumption?
3. When can we scale the test together and report a composite score (e.g. the ACT model), when does each subscale need to be considered separately (e.g. the SAT model)

In light of the results presented in the last chapter the following conclusions can be drawn:

1a. This hypothesis is generally correct – as the correlations increase the distinctions between the dimensions become difficult to identify. This is illustrated by the lower errors for higher correlations. For some settings the errors were larger for the 0.9 than for 0.95 correlations, indicating that correlation of 0.9 might be large enough for the subscales to become undistinguishable. However, it is surprising how small a role correlations played when errors were modeled with the glm compared to other factors. One can only conclude that unidimensional true score equating is relatively robust to low correlations between dimensions.

1b. This hypothesis was correct – more divergence in the two forms of the test makes tests less equatable. Since for most settings, information about the true ability is established from form 1 (because form 2 is put on the form 1 scale), any divergence from this form makes this information less valid. The drop for the 40-0 setting for no difference should be investigated closer, assuming it is not simulation variability. One of the hypotheses (excluding simulation variability) is that the last two settings are just too extreme – equating does not work even for the 20-20:40-0 setting, adding the 0-40:40-0 just adds noise, rather than information, to equating errors. A very careful examination of all the variables not included in this study (item discrimination and form information for example) would be necessary to unequivocally confirm this hypothesis. This condition is also partially confounded with the correlation between traits condition – removing five or even 20 items from one subscale if the correlations between traits are adequate mitigates the effect of non-compatible forms to be equated. It seems that the “subscale flip” (i.e.

replacing a form consisting of items of one dimension with a form consisting of items from another dimension) is not going to work for unidimensional equating – i.e. one cannot completely replace one subscale on one form with another subscale on the other form even if the correlations are very high. The glm analyses indicated that this is the main determining factor in predicting the bias and root mean squared difference error.

1c. This hypothesis was also correct – the more divergent the subpopulations the more difficult the equating. In particular lower ability examinees of both subpopulations were affected by the increase in the ability for one of the subpopulations. This increase was more pronounced with mean difference of 1 (one standard deviation) than of half a standard deviation. This was the only other factor (after divergence between forms) that was important in predicting the magnitude of all the error types examined. Given the design of this study it is understandable that the more able half of the population would have actually been penalized with equating – equating uses information about the examinees from form 1. This form, by design, in a sense disadvantages the examinees more able on the 1st dimension, because it gives fewer items that those examinees are good at than form 2 (which in turn advantages those examinees).

2. The answer to this question depends on the purpose(s) of the test. The standard error of measurement (SEM) can give some information on how aberrant the magnitudes of the errors discussed in previous chapters are. SEM is given by

$$SEM = \sigma * \sqrt{(1-\alpha)}$$

where σ is the standard deviation of the observed scores on the test and α is the reliability of the test. For the tests examined, the SEM for any setting was around 2.6. With this in mind, one can conclude that if only an overall population mean is desired, unidimensional true score equating of the multidimensional form is fairly robust to violations of unidimensionality. Regardless of the population ability distribution(s), even with lower correlations between traits (0.7 or even 0.5), small divergences (up to 5 items) are acceptable; with higher correlations, 20 items does not result in very large errors. This indicates that, at least with the employed here methodology, equating even with such a low correlation is reasonable (of course assuming that equating with 0.9 correlation is reasonable). However, if individual scores are of interest, as is the case in many testing situations, the large errors for the low and high ability students preclude anything but the parallel forms for anything but the highest (0.9 or 0.95) correlations where up to five items can differ in subscale from one form to the other. Examinees at the extremes of the ability scales are most effected by unidimensional equating of multidimensional trait. However, examines at the top and bottom of the ability continuum are not effected the same way by true score equating. As one subpopulation becomes more able and the mean ability for the whole population shifts, the errors for the top ability examines decrease – the estimation becomes less volatile with more people, leading to a decrease in errors. Conversely, the errors increase for the examinees at the low end of the ability scale. In fact, the sometimes counterintuitive results for the last setting and bottoms of each of the abilities and combined population might be due to a very small number of examinees at the top and bottom 10% of both abilities. The number of people (for the last simulation run) is shown in Tables in Appendix C. As can be seen the variability is quite large here,

but even 300 examinees constitute only 7% of the full population of 4000. Those results also clearly indicate a shift in the number of examinees used for estimation as the population mean increases. The ability estimation is naturally more precise in ranges where there are more examinees.

Moreover, this study unambiguously pinpoints the direction of the bias for the extremes of the population – equating underestimates the true score for top abilities of the dimension the test becomes heavier in (here θ_1) and overestimates for the bottom ability for this dimension. The reverse is true for the dimension the test becomes lighter in (θ_2).
3. While a correlation of around 0.7 (the general reported correlation between the (old) SAT sections) seems large enough to allow same construct equating, given that more than the overall population score is of interest, each subscale should be considered separately. Additionally, the difference between some subpopulations might be more than 1 on the standard normal scale. Differences larger than one were not examined in this study – errors would probably become more influenced by the relative differences in subpopulation ability as those differences get larger.

5.2 Improvements to methodology

While the M3PL of Reckase has been firmly entrenched in literature and research, it's appropriateness to the situation investigated here might be called into question. In particular, the difficulty parameter is assumed to be constant between dimensions, but intuitively, for a multidimensional case (especially when the traits are not highly correlated); the difficulty might be in a totally different metric.

The estimation of abilities, while confirmed by the pilot simulation study, is only approximate. It's unclear how well it works when the abilities have different means or follow completely different distributions. Should extensions of this study be desired, another approximation would be necessary for the two subpopulations differing in shape of their ability distribution. The difference in the magnitudes (although not the general trends) between the errors using calculated reference composite and estimated reference composite actually further underscores this. Alternatively, separate calculations for each of the subpopulation halves could have been employed. As mentioned in the results section, the exact reason for discrepancy between the errors using calculated and estimated reference composite is unknown and should be investigated further. It could be due to linearity of the calculated approximation. Another explanation is that the Parscale estimation takes into account the response data, while the calculation is purely based on both abilities and is linear.

This study underutilizes the generated item parameters which serve only to generate the observed scores. Those parameters could be used to calculate a more exact true score. It is unclear at this point if that would necessitate the use of the generated, rather than estimated item parameters for equating. One could employ the conditional standard error of measurement (standard deviation of the observed scores conditional on the true score) before and after the equating for an additional measure on how the multidimensional score translates into a unidimensional one.

Some of the settings in the current study might be called into question given the purpose of this study. For example, with a correlation of 0.5 between traits, equating should probably not be conducted for the whole form; rather, each of the traits should be

equated separately. While the results indicate equating functioning well for low correlations between traits, the validity of putting what are presumably different constructs on the same test form is questionable. In such a situation in practice, two constructs should (and hopefully would) be equated separately.

A unidimensional approximation was used for the calculation of the indices for the top and bottom abilities. This is very crude, and forces the error estimation to be based on very few people. A more population-dependent (maybe a rank-order) measure could have been employed.

Finally, the estimation of the item parameters depends on what Parscale is asked to do. However, given that the same estimation was used for each of the simulations it has little effect on the relative magnitudes of the equating indices. Population halves could have been estimated separately though for a more accurate ability estimation. However, one could argue, that in the real situation the person doing the equating would not know that a part of the population exhibits higher ability on one of the dimensions and population invariance requirement states that the equating functions have to be the same regardless of ability differences of subpopulations.

5.3 Further studies

This dissertation opens doors for a multitude of possible further studies. Naturally, the settings presented here can be extended for a fine-tuning of the results. Additional correlations of 0.6 and 0.8 could be considered in addition to the correlation of 1 which would serve as another baseline for the magnitude of the results. Lower correlations could also be explored to test the limits of robustness of IRT true-score

equating. Lower correlations however, could only serve the above purpose, since it's unlikely any real assessment would be equated together for low correlations between traits. Additional splits of the forms could also be explored, especially ones that are not as extreme – while the 20 items removed does not provide good equating, examining up to 10 items removed might shed some light on the unidimensional true score equating. In addition, that would allow for examining informative contrasts in the glm procedure – an estimate of the error increase with each item removed from a subscale could be obtained for instance. More ability differences could be examined, with varying sizes of groups differing in the mean ability. An interesting setting would be to have half of the population increase on one ability while the other half decreases on that ability to create a seeming "no difference in ability" scenario. Additional dimensions could be added, especially for higher correlations so that different dimensional structures of, for example, a mathematics assessment could be explored. Similarly, a lower correlational structure could be examined with more dimensions (correlations around 0.7 as in the reading constructs). The simulation study could be further enhanced if the correlations between dimensions were allowed to vary; however, care needs to be taken not to make the design so complicated it becomes a) not applicable to the real world and b) extremely difficult to interpret.

Single group design, while very intuitive and easy to interpret, is rarely used in operational settings. The reasons were described in detail in chapter 2.3.1. In the next iterations of this project, other designs could be used; for instance the common item (NEAT) design where a portion of items (an anchor) is common to both assessments. For a fixed percentage of common items some structures might be more robust to item

loading shift than others. Of course, it's not clear how many items are needed as the common items; there is a belief in the field that if the items are "good" (definitions of "good" vary) very few are needed. In general, the correlation between the score on the anchor and the score on the forms to be equated has to be high. Intuitively, it seems that a simple structure with common items would be robust to the variability in subscales since the correlation between the anchor score and the score on the items of the same dimension as the anchor would be high for each of the parts. It is also known, that common items have to reflect the forms that are to be equated (Kolen & Brennan, p. 19). This requirement obviously negates the purpose of this study in which the forms are changing. However, how close do multidimensional common items have to be to reflect both forms is an interesting research question. In the extreme case for instance – could one have common items reflecting the general Θ_{TT} and simple structure multidimensional test forms (which would reflect the general Θ_{TT})? Or will simple structure common items be able to provide good equating of non-simple structure forms? The NEAT design would also allow a setting in which part (or all) of the population increases in ability either on one dimension or both dimensions.

Another extension of the current study would be to examine a non-simple structure. However, with this kind of structure any potential results might be very difficult to interpret. And, as some argue, that such a structure exists only theoretically as a matter of test development agreement, rather than practically (although that could be caused simply by a lack of developed estimation procedures).

Another interesting setting which could be explored is not to keep the total number of items constant, but rather to remove items from one subscale without replacing

them in the other subscale. However, this course of action would cause reliability problems due to varying number of items. Same reliability of the forms, as discussed in detail previously (chapter 2.3), is one of the necessary conditions for equating to be valid. This problem could be partially overcome by increasing the discrimination of the remaining items, although that solution would be available only in a simulation study since it would presumably be hard to develop items of a given discrimination (unless the item pool is very rich). Ackerman (1992) warned against confounding form difficulty with equating, but it seems that some of this confounding is present in this study with the changing of the mean ability of the ½ of the population on the first dimension and changing of the items loading on each of the dimensions. As the forms diverge, the form being equated effectively becomes easier for half of the population (impacting the whole population in the process of course). The forms might be non-equatable because of that. However, purposefully confounding item difficulty and form could make for a very interesting further study. Closer monitoring of the quality of the items (in terms of item information) could not only help with the above issues, but in of itself provide interesting feedback on sensitivity of IRT true score equating to multidimensionality.

A linearity of the errors as a function of the factors was implicitly assumed in the generalized linear model procedure. It is conceivable that the errors are dependent on the higher degrees of the factors. More settings of form divergence and correlations could lead to an exploration of this issue.

Employing multiple equatings with each form differing very slightly from the previous form could give rise to interesting studies on the accumulation of errors as a multidimensional test is treated unidimensionally for many administrations.

References

Ackerman, T. (1994). Using multidimensional item response theory to understand what the items and tests are measuring. *Applied Measurement in Education*, 7, 255-278.

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, 20, 311-329.

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.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality of IRT true-score equating. *Applied Measurement in Education*, 12, 383-407.

Braun, H. I., & Holland. P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland and D.B. Rubin (Eds.), Test equating (p. 9-49). New York: Academic.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality of equating the law school admission test. *Journal of Educational Measurement*, 32, 79-96.

de Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement*, 33, 181-201.

Davey, T. , Oshima, T. C. , & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement,* 20, 405-416.

von Davier, A. A., Holland, P. W., & Thayer, D. (2004). The chain and post-stratification methods for observed-score equating: their relationship to population invariance. *Journal of Educational Measurement*, 2004, 41, 15-32.

Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel Haenszel procedure for detecting DIF. *Journal of Educational Statistics*, 18, 131-154.

Dorans, N.J. (1954). *Approximate IRT formula score and scaled score standard errors of measurement at different ability levels* (Rep. No. SR-84-118). Princeton, NJ: Educational Testing Service.

Dorans, N. J. (2000). Distinctions among classes of linkages. *The College Board Research Notes*, RN-11.

Dorans, N. J. (2004c). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, 2004, 41, 43-68.

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.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, 14, 31-57.

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true- and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, 10, 105-121.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation of the common-item equating design. *Applied Psychological Measurement*, 26, 3-24.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6, 195-240.

Holland, P., & Dorans, N. (2006). Linking and Equating. In R. L. Brennan (Ed.), *Educational Measurement* (pp 187-220). Westport: National Council on Measurement in Education and American Council on Education.

Kim, S., & Cohen A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22, 131-143.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, 43, 53-76.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking. Methods and practices*. New York: Springer-Verlag.

Liu, J., Cahn, M. F., & Dorans, N. (2006). An application of score equity assessment: invariance of linkage of New SAT to Old SAT across gender groups. *Journal of Educational Measurement*, 43, 113-129.

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.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

te Marvelde, J. M., Glas, C. A., van Landeghem, G., & van Damme, J. (2006). Application of multidimensional item response theory models to longitudinal data. *Educational and Psychological Measurement*, 66, 5-34

Morris, C. N. (1982). On the foundations of test equating. In P.W. Holland and D.B. Rubin (Eds.), *Test equating* (p. 169-191). New York: Academic.

Muraki, E., & Bock, R. D. (1993). PARSCALEL IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. Chicago, IL: Scientific Software International.

Oshima, T. C, Davey, T. C., & Lee, K. (2000). Multidimensional linking: four practical approaches. *Journal of Educational Measurement*, 37, 357-374.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.

Reckase, M., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, 15, 361-373.

Ree, M. J., Carretta, T. R., & Earles, J. A., (2003). Salvaging construct equivalence though equating. *Personality and Individual Differences*, 35, 1293-1305.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617.

Stout, W. (1990). A new response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293-352.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement*, 29, 418-432.

Wang, M. M. (1985). Fitting a unidimenisional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT. Unpublished doctoral dissertation. University of Iowa. Iowa City.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2003). TESTFACT 4L Test scoring, items statistics, and item factor analysis. Mooresville, IN: Scientific Software.

Zhang, J. (2004), Conditional covariance theory and DETECT for polytomous items (Rep No. RR-04-50). Princeton, NJ: Educational Testing Service.

Zhang, J., Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, 64, 129-152.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG. Multiple-group IRT analysis and test maintenance for binary items.* Chicago, IL: SSI Scientific Software International.

APPENDIX A

Full results of the simulation study.

The following pages list the errors (MSD and RMSD) for the two models examined in this simulation study by the distribution of items on the second form (listed in the first column, uniquely identified for all but the most divergent setting) and the correlations between simulated traits. Results using the calculated reference composite for the 3PL model start below and continue through Table A54 on page 111. Results using the calculated reference composite for the 1PL model start with Table A55 on page 111 and continue through Table A108 on page 129. Results using the estimated reference composite for the 3PL model start with Table A109 on page 130 and continue through Table A162 on page 147. Results using the estimated reference composite for the 1PL model start with Table A163 on page 148 and continue though Table A214. For each of the sections Mean Signed Difference is listed first for all the population sections examined, followed by the Root Mean Squared Difference.

Table A1: MSD using calculated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.000	-0.001	0.000
25-15	-0.004	0.012	0.001	0.003
40-0	-0.045	0.008	-0.005	0.033
0-40:40-0	0.038	-0.037	-0.007	-0.001

Table A2: MSD using calculated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.005	-0.003	-0.003	-0.003
25-15	-0.116	-0.096	-0.087	-0.063
40-0	-1.496	-1.274	-0.997	-0.876
0-40:40-0	-0.076	-0.014	0.004	0.046

Table A3: MSD using calculated reference composite top 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.224	0.159	0.086	0.057
25-15	0.924	0.699	0.359	0.202
40-0	2.153	1.256	0.336	-0.013
0-40:40-0	8.258	6.189	3.535	2.286

Table A4: MSD using calculated reference composite top 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.237	-0.169	-0.098	-0.063
25-15	-1.260	-0.955	-0.556	-0.365
40-0	-6.731	-4.950	-3.042	-2.194
0-40:40-0	-8.119	-6.014	-3.600	-2.240

Table A5: MSD using calculated reference composite for bottom 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.003	0.003	0.003	0.003
25-15	0.112	0.096	0.068	0.069
40-0	1.488	1.243	0.951	0.946
0-40:40-0	-0.102	-0.033	0.002	0.004

Table A6: MSD using calculated reference composite for the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.197	-0.152	-0.079	-0.055
25-15	-0.901	-0.633	-0.309	-0.199
40-0	-1.941	-1.240	-0.309	0.040
0-40:40-0	-8.412	-6.317	-3.156	-2.382

Table A7: MSD using calculated reference composite for the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.215	0.166	0.087	0.059
25-15	1.210	0.927	0.495	0.383
40-0	6.393	4.924	2.993	2.392
0-40:40-0	8.376	6.207	3.107	2.440

Table A8: MSD using calculated reference composite for subpopulation 1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.001	0.001	0.000	0.000
25-15	0.004	0.020	0.004	0.006
40-0	-0.030	0.023	0.001	0.026
0-40:40-0	0.054	-0.021	0.011	0.009

Table A9: MSD using calculated reference composite for subpopulation 2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.000	-0.001	0.001
25-15	-0.011	0.003	-0.001	0.001
40-0	-0.060	-0.006	-0.010	0.040
0-40:40-0	0.022	-0.053	-0.026	-0.012

Table A10: RMSD using calculated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.199	0.161	0.092	0.067
25-15	0.974	0.765	0.460	0.333
40-0	4.055	3.141	1.972	1.462
0-40:40-0	7.166	5.899	3.356	2.369

Table A11: RMSD using calculated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.044	0.039	0.032	0.022
25-15	0.258	0.223	0.181	0.125
40-0	1.796	1.553	1.200	1.011
0-40:40-0	1.939	1.570	1.393	0.982

Table A12: RMSD using calculated reference composite top 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.273	0.191	0.101	0.066
25-15	1.170	0.876	0.446	0.258
40-0	3.341	2.074	0.872	0.543
0-40:40-0	9.591	7.298	4.053	2.620

Table A13: RMSD using calculated reference composite top 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.283	0.201	0.113	0.071
25-15	1.456	1.092	0.618	0.402
40-0	7.278	5.418	3.252	2.286
0-40:40-0	9.443	7.107	4.130	2.571

Table A14: RMSD using calculated reference composite for bottom 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.041	0.037	0.027	0.022
25-15	0.230	0.227	0.148	0.137
40-0	1.781	1.526	1.143	1.082
0-40:40-0	1.934	1.748	1.220	0.976

Table A15: RMSD using calculated reference composite for the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.242	0.186	0.093	0.064
25-15	1.143	0.788	0.390	0.250
40-0	3.019	2.035	0.870	0.559
0-40:40-0	9.695	7.434	3.643	2.723

Table A16: RMSD using calculated reference composite for the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.260	0.200	0.102	0.067
25-15	1.407	1.072	0.557	0.418
40-0	6.971	5.386	3.210	2.520
0-40:40-0	9.684	7.350	3.591	2.795

Table A17: RMSD using calculated reference composite for subpopulation 1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.199	0.160	0.093	0.067
25-15	0.971	0.765	0.459	0.334
40-0	4.046	3.134	1.966	1.463
0-40:40-0	7.118	5.927	3.379	2.367

Table A18: RMSD using calculated reference composite for subpopulation 2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.199	0.161	0.092	0.067
25-15	0.977	0.766	0.460	0.331
40-0	4.063	3.148	1.977	1.460
0-40:40-0	7.214	5.871	3.333	2.370

Table A19: MSD using calculated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.047	0.049	0.048	0.048
25-15	0.215	0.235	0.236	0.234
40-0	0.760	0.806	0.820	0.855
0-40:40-0	1.746	1.814	1.799	1.829

Table A20: MSD using calculated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.004	0.008	0.010
25-15	-0.069	-0.040	-0.008	0.002
40-0	-0.843	-0.816	-0.571	-0.421
0-40:40-0	0.183	0.226	0.374	0.453

Table A21: MSD using calculated reference composite top 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.206	0.165	0.094	0.073
25-15	0.875	0.637	0.343	0.270
40-0	1.529	1.208	0.479	0.267
0-40:40-0	8.277	5.921	3.625	2.883

Table A22: MSD using calculated reference composite top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.188	-0.130	-0.063	-0.045
25-15	-1.034	-0.704	-0.369	-0.282
40-0	-5.433	-4.431	-2.302	-1.763
0-40:40-0	-6.630	-4.502	-2.361	-1.578

Table A23: MSD using calculated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.010	0.012	0.017	0.017
25-15	0.160	0.148	0.135	0.151
40-0	2.068	1.620	1.490	1.635
0-40:40-0	0.253	0.431	0.542	0.646

Table A24: MSD using calculated reference composite for the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.222	-0.165	-0.097	-0.068
25-15	-1.041	-0.760	-0.407	-0.294
40-0	-2.577	-1.625	-0.643	-0.198
0-40:40-0	-9.224	-7.144	-4.075	-2.786

Table A25: MSD using calculated reference composite for the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.282	0.219	0.147	0.115
25-15	1.552	1.262	0.824	0.708
40-0	7.276	5.967	4.198	3.789
0-40:40-0	10.628	8.665	5.652	4.501

Table A26: MSD using calculated reference composite for subpopulation 1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.002	0.000	0.001	-0.001
25-15	-0.012	0.008	0.002	0.004
40-0	0.126	0.094	0.072	0.076
0-40:40-0	-0.026	-0.020	-0.021	0.016

Table A27: MSD using calculated reference composite for subpopulation 2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.096	0.098	0.096	0.096
25-15	0.442	0.462	0.470	0.463
40-0	1.394	1.519	1.568	1.634
0-40:40-0	3.517	3.647	3.619	3.641

Table A28: RMSD using calculated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.224	0.175	0.116	0.098
25-15	1.053	0.852	0.590	0.479
40-0	4.138	3.420	2.252	1.943
0-40:40-0	7.754	6.462	4.346	3.640

Table A29: RMSD using calculated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.025	0.029	0.023	0.021
25-15	0.167	0.146	0.100	0.089
40-0	1.038	1.035	0.760	0.559
0-40:40-0	1.311	0.980	0.983	0.879

Table A30: RMSD using calculated reference composite top 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.268	0.203	0.111	0.084
25-15	1.129	0.819	0.425	0.325
40-0	2.517	2.020	0.907	0.602
0-40:40-0	9.776	7.185	4.192	3.286

Table A31: RMSD using calculated reference composite top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.250	0.159	0.073	0.052
25-15	1.261	0.833	0.428	0.311
40-0	6.205	5.000	2.469	1.836
0-40:40-0	8.056	5.530	2.866	1.836

Table A32: RMSD using calculated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.052	0.048	0.047	0.039
25-15	0.319	0.302	0.253	0.252
40-0	2.537	2.012	1.798	1.912
0-40:40-0	2.686	2.561	1.859	1.737

Table A33: RMSD using calculated reference composite for the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.277	0.202	0.113	0.079
25-15	1.287	0.935	0.499	0.361
40-0	3.806	2.495	1.230	0.752
0-40:40-0	10.445	8.186	4.649	3.192

Table A34: RMSD using calculated reference composite for the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.340	0.258	0.166	0.129
25-15	1.764	1.432	0.928	0.777
40-0	7.728	6.425	4.488	3.971
0-40:40-0	11.903	9.857	6.375	5.037

Table A35: RMSD using calculated reference composite for subpopulation 1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.215	0.160	0.092	0.066
25-15	1.007	0.784	0.474	0.327
40-0	4.088	3.264	1.910	1.473
0-40:40-0	7.414	5.961	3.490	2.488

Table A36: RMSD using calculated reference composite for subpopulation 2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.233	0.188	0.136	0.121
25-15	1.097	0.914	0.687	0.592
40-0	4.186	3.569	2.548	2.319
0-40:40-0	8.078	6.926	5.060	4.507

Table A37: MSD using calculated reference composite full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.089	0.090	0.087	0.088
25-15	0.403	0.402	0.435	0.417
40-0	1.241	1.388	1.471	1.528
0-40:40-0	3.515	3.371	3.316	3.388

Table A38: MSD using calculated reference composite top 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.005	0.007	0.010	0.011
25-15	-0.031	-0.005	0.022	0.026
40-0	-0.440	-0.307	-0.187	-0.170
0-40:40-0	0.248	0.277	0.390	0.388

Table A39: MSD using calculated reference composite top 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.215	0.169	0.096	0.088
25-15	0.796	0.623	0.455	0.340
40-0	1.355	0.977	0.623	0.490
0-40:40-0	8.310	5.731	3.608	3.160

Table A40: MSD using calculated reference composite top 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.144	-0.101	-0.044	-0.034
25-15	-0.813	-0.550	-0.310	-0.198
40-0	-4.487	-2.937	-1.736	-1.141
0-40:40-0	-5.578	-3.092	-1.383	-1.019

Table A41: MSD using calculated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.015	0.016	0.021	0.025
25-15	0.193	0.190	0.175	0.199
40-0	2.408	2.142	1.840	1.808
0-40:40-0	0.349	0.646	0.806	0.804

Table A42: MSD using calculated reference composite for the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.232	-0.199	-0.126	-0.085
25-15	-1.165	-0.913	-0.516	-0.372
40-0	-3.342	-2.554	-1.160	-0.588
0-40:40-0	-10.374	-8.430	-5.222	-3.760

Table A43: MSD using calculated reference composite for the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.337	0.308	0.247	0.232
25-15	1.855	1.594	1.279	1.219
40-0	7.661	7.085	5.750	5.821
0-40:40-0	13.732	12.009	9.666	9.167

Table A44: MSD using calculated reference composite for subpopulation 1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.000	-0.001	0.000
25-15	0.012	-0.006	0.012	0.010
40-0	0.017	0.108	0.153	0.150
0-40:40-0	-0.020	-0.011	0.025	0.009

Table A45: MSD using calculated reference composite for subpopulation 2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.176	0.180	0.174	0.175
25-15	0.794	0.809	0.859	0.823
40-0	2.465	2.669	2.789	2.907
0-40:40-0	7.051	6.753	6.607	6.766

Table A46: RMSD using calculated reference composite full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.233	0.206	0.159	0.147
25-15	1.142	0.970	0.794	0.711
40-0	4.372	3.812	2.961	2.848
0-40:40-0	9.240	7.690	6.075	5.714

Table A47: RMSD using calculated reference composite top 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.020	0.018	0.017	0.019
25-15	0.099	0.088	0.087	0.068
40-0	0.555	0.425	0.319	0.291
0-40:40-0	0.624	0.552	0.609	0.608

Table A48: RMSD using calculated reference composite top 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.278	0.216	0.116	0.100
25-15	1.051	0.812	0.542	0.395
40-0	2.244	1.577	0.924	0.680
0-40:40-0	10.258	7.204	4.349	3.650

Table A49: RMSD using calculated reference composite top 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.192	0.129	0.053	0.038
25-15	1.054	0.678	0.349	0.216
40-0	5.500	3.598	1.912	1.196
0-40:40-0	7.417	4.118	1.687	1.220

Table A50: RMSD using calculated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.077	0.070	0.071	0.067
25-15	0.424	0.416	0.365	0.382
40-0	3.106	2.806	2.408	2.321
0-40:40-0	3.503	3.527	3.045	2.718

Table A51: RMSD using calculated reference composite for the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.271	0.229	0.143	0.096
25-15	1.387	1.078	0.604	0.443
40-0	4.485	3.440	1.711	1.103
0-40:40-0	11.644	9.403	5.881	4.224

Table A52: RMSD using calculated reference composite for the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.383	0.346	0.272	0.251
25-15	2.032	1.756	1.398	1.301
40-0	8.119	7.511	6.064	6.095
0-40:40-0	14.972	13.093	10.519	9.848

Table A53: RMSD using calculated reference composite for subpopulation 1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.196	0.159	0.090	0.063
25-15	1.016	0.775	0.460	0.321
40-0	4.168	3.333	1.973	1.538
0-40:40-0	8.077	6.138	3.598	2.567

Table A54: RMSD using calculated reference composite for subpopulation 2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.264	0.244	0.206	0.198
25-15	1.256	1.131	1.024	0.952
40-0	4.566	4.237	3.692	3.722
0-40:40-0	10.271	8.977	7.801	7.662

Table A55: MSD using calculated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.001	0.001	0.002	0.001
25-15	0.006	0.004	0.010	0.012
40-0	0.057	0.076	0.139	0.179
0-40:40-0	0.026	-0.010	0.036	0.010

Table A56: MSD using calculated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.006	-0.004	-0.006	-0.003
25-15	-0.119	-0.099	-0.071	-0.068
40-0	-1.598	-1.335	-1.044	-0.819
0-40:40-0	0.060	-0.002	-0.027	0.003

Table A57: MSD using calculated reference composite top 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.204	0.150	0.070	0.049
25-15	0.921	0.634	0.327	0.192
40-0	1.942	1.259	0.303	-0.040
0-40:40-0	8.050	5.857	3.079	2.243

Table A58: MSD using calculated reference composite top 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.219	-0.160	-0.086	-0.058
25-15	-1.235	-0.885	-0.517	-0.348
40-0	-6.143	-4.797	-2.770	-2.078
0-40:40-0	-8.068	-5.815	-3.128	-2.132

Table A59: MSD using calculated reference composite for bottom 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.004	0.006	0.013	0.006
25-15	0.198	0.191	0.126	0.124
40-0	2.703	2.278	1.884	1.884
0-40:40-0	-0.047	-0.027	0.013	-0.005

Table A60: MSD using calculated reference composite for the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.288	-0.216	-0.120	-0.080
25-15	-1.248	-0.940	-0.475	-0.328
40-0	-2.694	-1.803	-0.494	0.085
0-40:40-0	-10.376	-8.343	-4.627	-3.182

Table A61: MSD using calculated reference composite for the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.305	0.232	0.150	0.090
25-15	1.604	1.240	0.751	0.561
40-0	7.405	6.099	3.994	3.533
0-40:40-0	10.292	8.355	4.700	3.208

Table A62: MSD using calculated reference composite for subpopulation 1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.001	0.000	0.003	0.000
25-15	0.006	0.000	0.010	0.008
40-0	0.058	0.058	0.139	0.193
0-40:40-0	-0.006	0.024	0.018	-0.016

Table A63: MSD using calculated reference composite for subpopulation 2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.001	0.001	0.001
25-15	0.007	0.009	0.010	0.016
40-0	0.055	0.095	0.139	0.165
0-40:40-0	0.059	-0.045	0.054	0.036

Table A64: RMSD using calculated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.218	0.171	0.114	0.069
25-15	1.061	0.836	0.496	0.368
40-0	4.133	3.324	1.999	1.653
0-40:40-0	7.727	6.135	3.616	2.499

Table A65: RMSD using calculated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.042	0.039	0.024	0.023
25-15	0.257	0.220	0.169	0.130
40-0	1.902	1.633	1.231	0.941
0-40:40-0	1.926	1.722	1.174	1.027

Table A66: RMSD using calculated reference composite top 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.248	0.183	0.086	0.057
25-15	1.159	0.804	0.408	0.242
40-0	3.013	2.056	0.820	0.479
0-40:40-0	9.411	6.866	3.561	2.545

Table A67: RMSD using calculated reference composite top 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.264	0.191	0.100	0.066
25-15	1.429	1.031	0.580	0.383
40-0	6.734	5.247	2.947	2.176
0-40:40-0	9.435	6.854	3.576	2.453

Table A68: RMSD using calculated reference composite for bottom 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.086	0.078	0.054	0.045
25-15	0.477	0.445	0.300	0.247
40-0	3.171	2.750	2.208	2.117
0-40:40-0	3.296	3.341	2.264	1.768

Table A69: RMSD using calculated reference composite for the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.334	0.250	0.142	0.091
25-15	1.486	1.135	0.575	0.404
40-0	4.084	2.866	1.283	0.918
0-40:40-0	11.659	9.427	5.257	3.558

Table A70: RMSD using calculated reference composite for the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.346	0.265	0.168	0.101
25-15	1.780	1.374	0.822	0.611
40-0	7.798	6.448	4.183	3.648
0-40:40-0	11.551	9.441	5.283	3.595

Table A71: RMSD using calculated reference composite for subpopulation 1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.218	0.171	0.114	0.068
25-15	1.064	0.837	0.497	0.367
40-0	4.133	3.315	1.987	1.660
0-40:40-0	7.698	6.147	3.622	2.503

Table A72: RMSD using calculated reference composite for subpopulation 2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.217	0.171	0.114	0.069
25-15	1.059	0.835	0.496	0.369
40-0	4.133	3.333	2.011	1.646
0-40:40-0	7.756	6.122	3.609	2.495

Table A73: MSD using calculated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.051	0.052	0.050	0.053
25-15	0.250	0.258	0.261	0.257
40-0	0.810	0.901	1.078	1.023
0-40:40-0	1.839	1.820	1.859	1.889

Table A74: MSD using calculated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.004	0.007	0.011
25-15	-0.064	-0.039	-0.009	0.004
40-0	-0.960	-0.697	-0.415	-0.422
0-40:40-0	0.218	0.228	0.456	0.490

Table A75: MSD using calculated reference composite top 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.200	0.152	0.082	0.069
25-15	0.921	0.639	0.344	0.271
40-0	1.625	1.044	0.406	0.243
0-40:40-0	7.613	5.366	3.309	2.682

Table A76: MSD using calculated reference composite top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.176	-0.123	-0.059	-0.045
25-15	-1.047	-0.712	-0.363	-0.266
40-0	-5.302	-3.714	-1.888	-1.666
0-40:40-0	-6.293	-4.129	-2.183	-1.537

Table A77: MSD using calculated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.023	0.021	0.027	0.029
25-15	0.276	0.279	0.254	0.263
40-0	3.237	2.913	2.568	2.516
0-40:40-0	0.472	0.682	0.972	1.030

Table A78: MSD using calculated reference composite for the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.290	-0.218	-0.137	-0.091
25-15	-1.328	-1.007	-0.539	-0.360
40-0	-3.451	-2.500	-0.957	-0.240
0-40:40-0	-11.303	-8.784	-5.047	-3.923

Table A79: MSD using calculated reference composite for the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.350	0.281	0.197	0.157
25-15	1.891	1.579	1.067	0.879
40-0	8.206	7.159	5.618	4.629
0-40:40-0	12.900	10.747	7.033	5.854

Table A80: MSD using calculated reference composite for subpopulation 1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	0.000	-0.001	0.001
25-15	0.010	0.017	0.011	0.013
40-0	0.118	0.154	0.256	0.215
0-40:40-0	0.007	0.000	0.038	-0.001

Table A81: MSD using calculated reference composite for subpopulation 2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.100	0.103	0.102	0.104
25-15	0.490	0.498	0.511	0.502
40-0	1.502	1.648	1.900	1.831
0-40:40-0	3.670	3.640	3.679	3.780

Table A82: RMSD using calculated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.226	0.185	0.125	0.106
25-15	1.160	0.942	0.637	0.516
40-0	4.410	3.627	2.699	2.208
0-40:40-0	8.391	6.669	4.435	3.792

Table A83: RMSD using calculated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.029	0.029	0.022	0.022
25-15	0.164	0.142	0.099	0.092
40-0	1.148	0.876	0.553	0.552
0-40:40-0	1.171	0.999	0.987	0.902

Table A84: RMSD using calculated reference composite top 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.249	0.187	0.097	0.080
25-15	1.197	0.814	0.428	0.324
40-0	2.578	1.685	0.760	0.558
0-40:40-0	9.267	6.600	3.841	3.009

Table A85: RMSD using calculated reference composite top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.221	0.151	0.070	0.051
25-15	1.270	0.854	0.416	0.290
40-0	6.056	4.318	2.056	1.752
0-40:40-0	7.907	5.172	2.569	1.780

Table A86: RMSD using calculated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.106	0.095	0.084	0.067
25-15	0.594	0.560	0.459	0.420
40-0	3.941	3.598	3.073	2.885
0-40:40-0	4.540	4.282	3.212	2.851

Table A87: RMSD using calculated reference composite for the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.333	0.251	0.159	0.106
25-15	1.579	1.195	0.655	0.437
40-0	4.790	3.655	1.863	1.034
0-40:40-0	12.610	9.846	5.652	4.455

Table A88: RMSD using calculated reference composite for the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.390	0.315	0.220	0.173
25-15	2.075	1.732	1.166	0.947
40-0	8.628	7.540	5.890	4.820
0-40:40-0	14.181	11.836	7.793	6.433

Table A89: RMSD using calculated reference composite for subpopulation 1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.218	0.170	0.101	0.072
25-15	1.122	0.884	0.515	0.356
40-0	4.404	3.509	2.274	1.682
0-40:40-0	8.175	6.264	3.587	2.649

Table A90: RMSD using calculated reference composite for subpopulation 2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.233	0.199	0.145	0.131
25-15	1.196	0.997	0.739	0.637
40-0	4.416	3.740	3.066	2.631
0-40:40-0	8.599	7.050	5.145	4.663

Table A91: MSD using calculated reference composite full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.094	0.093	0.096	0.096
25-15	0.445	0.466	0.465	0.453
40-0	1.414	1.507	1.651	1.729
0-40:40-0	3.506	3.532	3.653	3.597

Table A92: MSD using calculated reference composite top 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.006	0.008	0.010	0.010
25-15	-0.020	-0.003	0.021	0.024
40-0	-0.374	-0.320	-0.163	-0.168
0-40:40-0	0.315	0.322	0.384	0.371

Table A93: MSD using calculated reference composite top 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.210	0.148	0.100	0.080
25-15	0.761	0.566	0.376	0.338
40-0	1.114	0.888	0.515	0.490
0-40:40-0	7.768	5.628	3.667	2.932

Table A94: MSD using calculated reference composite top 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.144	-0.092	-0.046	-0.031
25-15	-0.765	-0.499	-0.264	-0.188
40-0	-4.165	-2.844	-1.449	-1.136
0-40:40-0	-4.886	-3.073	-1.468	-0.817

Table A95: MSD using calculated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.026	0.030	0.037	0.037
25-15	0.309	0.308	0.299	0.294
40-0	3.910	3.309	2.919	2.793
0-40:40-0	0.512	0.974	1.192	1.373

Table A96: MSD using calculated reference composite for the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	-0.295	-0.246	-0.149	-0.107
25-15	-1.519	-1.224	-0.692	-0.455
40-0	-4.581	-3.377	-1.614	-0.727
0-40:40-0	-12.533	-9.858	-6.469	-4.732

Table A97: MSD using calculated reference composite for the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.397	0.363	0.291	0.283
25-15	2.195	1.972	1.598	1.385
40-0	8.806	7.779	6.931	6.555
0-40:40-0	15.517	13.559	11.578	11.094

Table A98: MSD using calculated reference composite for subpopulation 1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.002	-0.001	0.002	0.001
25-15	0.012	0.024	0.019	0.025
40-0	0.241	0.233	0.268	0.280
0-40:40-0	-0.051	0.055	-0.024	-0.009

Table A99: MSD using calculated reference composite for subpopulation 2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.185	0.187	0.190	0.191
25-15	0.878	0.907	0.912	0.880
40-0	2.588	2.780	3.035	3.178
0-40:40-0	7.063	7.010	7.329	7.202

Table A100: RMSD using calculated reference composite full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.250	0.220	0.175	0.164
25-15	1.292	1.124	0.880	0.770
40-0	4.934	4.119	3.393	3.189
0-40:40-0	9.753	8.200	6.829	6.243

Table A101: RMSD using calculated reference composite top 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.022	0.018	0.018	0.016
25-15	0.082	0.072	0.065	0.072
40-0	0.475	0.438	0.272	0.298
0-40:40-0	0.654	0.643	0.618	0.552

Table A102: RMSD using calculated reference composite top 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.265	0.189	0.119	0.092
25-15	1.031	0.757	0.466	0.394
40-0	1.842	1.428	0.773	0.678
0-40:40-0	9.807	7.065	4.444	3.402

Table A103: RMSD using calculated reference composite top 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.184	0.122	0.055	0.035
25-15	1.013	0.626	0.305	0.207
40-0	5.391	3.454	1.628	1.201
0-40:40-0	6.632	4.146	1.830	0.980

Table A104: RMSD using calculated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.131	0.131	0.107	0.101
25-15	0.773	0.731	0.624	0.558
40-0	4.968	4.320	3.797	3.556
0-40:40-0	5.605	5.455	4.477	4.318

Table A105: RMSD using calculated reference composite for the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.332	0.280	0.170	0.123
25-15	1.753	1.419	0.812	0.534
40-0	5.997	4.413	2.324	1.364
0-40:40-0	13.833	10.985	7.234	5.283

Table A106: RMSD using calculated reference composite for the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.435	0.399	0.317	0.304
25-15	2.375	2.141	1.717	1.475
40-0	9.365	8.225	7.273	6.849
0-40:40-0	16.773	14.669	12.529	11.831

Table A107: RMSD using calculated reference composite for subpopulation 1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.219	0.177	0.103	0.074
25-15	1.178	0.919	0.533	0.360
40-0	4.932	3.731	2.330	1.794
0-40:40-0	8.767	6.724	4.117	2.881

Table A108: RMSD using calculated reference composite for subpopulation 2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20				
21-19	0.278	0.256	0.226	0.220
25-15	1.397	1.296	1.125	1.027
40-0	4.934	4.472	4.194	4.138
0-40:40-0	10.646	9.448	8.735	8.345

Table A109: MSD using estimated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.010	0.082	0.105	-0.013
21-19	-0.052	-0.069	0.071	0.032
25-15	0.059	-0.019	-0.029	0.009
40-0	0.316	0.121	-0.033	0.097
0-40:40-0	0.039	-0.112	-0.026	0.055

Table A110: MSD using estimated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.261	-0.032	0.447	-0.006
21-19	-0.068	-0.053	0.020	0.124
25-15	0.203	-0.052	-0.051	-0.006
40-0	-0.751	-0.380	-0.230	-0.098
0-40:40-0	0.119	-0.049	0.024	0.006

Table A111: MSD using estimated reference composite top 10% of $\theta 1$, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.236	-0.020	0.346	0.030
21-19	0.644	0.193	0.113	0.203
25-15	1.665	0.773	0.370	0.229
40-0	3.529	2.643	1.353	1.024
0-40:40-0	8.010	5.856	3.116	2.221

Table A112: MSD using estimated reference composite top 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.396	-0.007	0.339	0.000
21-19	-0.756	-0.288	-0.077	0.070
25-15	-1.519	-0.975	-0.473	-0.265
40-0	-4.405	-3.326	-1.821	-1.113
0-40:40-0	-7.827	-6.032	-3.096	-2.172

Table A113: MSD using estimated reference composite for bottom 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.285	0.116	-0.425	-0.033
21-19	-0.073	-0.068	-0.033	-0.200
25-15	-0.221	0.042	0.017	-0.026
40-0	0.741	0.292	0.158	0.089
0-40:40-0	-0.105	0.004	-0.039	-0.015

Table A114: MSD using estimated reference composite for the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.420	0.139	-0.344	0.154
21-19	-0.150	-0.101	-0.043	-0.127
25-15	-1.123	-0.955	-0.420	-0.212
40-0	-2.227	-2.082	-1.197	-0.816
0-40:40-0	-7.554	-5.374	-2.955	-1.899

Table A115: MSD using estimated reference composite for the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.125	0.129	-0.260	-0.024
21-19	0.061	0.036	0.154	0.007
25-15	1.474	1.175	0.453	0.241
40-0	5.177	3.475	1.702	1.284
0-40:40-0	7.429	5.263	2.897	2.017

Table A116: MSD using calculated reference composite for subpopulation 1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.013	0.079	0.120	-0.016
21-19	-0.054	-0.076	0.068	0.030
25-15	0.060	-0.027	-0.029	0.004
40-0	0.308	0.119	-0.031	0.105
0-40:40-0	0.035	-0.118	-0.026	0.058

Table A117: MSD using calculated reference composite for subpopulation 2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.007	0.085	0.089	-0.010
21-19	-0.051	-0.061	0.074	0.034
25-15	0.058	-0.011	-0.029	0.014
40-0	0.324	0.123	-0.036	0.088
0-40:40-0	0.044	-0.106	-0.026	0.052

Table A118: RMSD using estimated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.343	3.151	3.358	3.113
21-19	3.263	3.095	3.092	3.132
25-15	3.579	3.220	3.080	3.064
40-0	4.935	4.324	3.488	3.297
0-40:40-0	7.520	6.204	4.379	3.778

Table A119: RMSD using estimated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.231	1.978	2.352	1.945
21-19	3.263	1.935	1.958	2.090
25-15	2.192	1.943	1.957	1.972
40-0	2.274	2.256	2.088	2.022
0-40:40-0	2.711	2.517	2.292	2.247

Table A120: RMSD using estimated reference composite top 10% of $\theta 1$, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.381	2.754	2.931	2.580
21-19	3.257	2.811	2.603	2.631
25-15	3.479	2.827	2.602	2.546
40-0	5.310	4.277	2.964	2.735
0-40:40-0	9.657	7.370	4.432	3.600

Table A121: RMSD using estimated reference composite top 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.262	2.787	2.975	2.450
21-19	3.288	2.877	2.648	2.673
25-15	3.811	3.079	2.713	2.575
40-0	5.766	4.849	3.389	2.930
0-40:40-0	9.576	7.515	4.384	3.570

Table A122: RMSD using estimated reference composite for bottom 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.335	2.475	2.954	2.444
21-19	2.088	2.217	2.289	2.467
25-15	2.403	2.241	2.295	2.319
40-0	2.451	2.555	2.312	2.384
0-40:40-0	2.799	2.548	2.462	2.462

Table A123: RMSD using estimated reference composite for the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.468	3.086	3.254	2.815
21-19	3.125	2.999	2.942	2.966
25-15	3.337	3.081	2.868	2.830
40-0	4.260	3.921	3.063	2.944
0-40:40-0	9.335	6.984	4.455	3.585

Table A124: RMSD using estimated reference composite for the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.286	3.131	3.457	2.803
21-19	3.267	3.055	2.925	2.943
25-15	3.997	3.439	2.938	2.895
40-0	6.733	5.232	3.534	3.276
0-40:40-0	9.284	6.847	4.396	3.651

Table A125: RMSD using calculated reference composite for subpopulation 1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.351	3.149	3.391	3.108
21-19	3.259	3.093	3.088	3.131
25-15	3.570	3.215	3.085	3.061
40-0	4.928	4.327	3.488	3.298
0-40:40-0	7.516	6.196	4.390	3.777

Table A126: RMSD using calculated reference composite for subpopulation 2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.336	3.152	3.323	3.117
21-19	3.266	3.097	3.095	3.134
25-15	3.587	3.224	3.076	3.067
40-0	4.940	4.321	3.487	3.295
0-40:40-0	7.523	6.211	4.367	3.780

Table A127: MSD using estimated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.117	-0.082	0.012	0.070
21-19	-0.123	-0.001	0.054	-0.062
25-15	0.261	0.076	0.152	0.056
40-0	0.576	0.335	0.151	0.175
0-40:40-0	0.360	0.263	0.342	0.215

Table A128: MSD using estimated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.343	-0.078	0.010	0.001
21-19	0.213	-0.007	-0.013	-0.019
25-15	0.199	-0.011	0.014	0.003
40-0	-0.472	-0.406	-0.236	-0.128
0-40:40-0	-0.004	0.199	0.025	0.029

Table A129: MSD using estimated reference composite top 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.515	-0.160	0.009	0.017
21-19	0.088	0.350	0.069	0.036
25-15	2.154	0.927	0.328	0.348
40-0	3.599	2.604	1.461	1.087
0-40:40-0	7.580	5.799	3.254	2.470

Table A130: MSD using estimated reference composite top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.077	-0.043	0.002	0.011
21-19	-0.074	-0.429	-0.032	-0.056
25-15	-1.979	-1.094	-0.229	-0.485
40-0	-4.084	-3.320	-1.905	-1.416
0-40:40-0	-7.237	-5.210	-2.943	-2.396

Table A131: MSD using estimated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.097	-0.029	0.017	-0.066
21-19	-0.187	-0.006	0.012	-0.029
25-15	0.145	0.102	-0.049	-0.013
40-0	0.951	0.591	0.285	0.188
0-40:40-0	-0.021	-0.240	0.002	-0.132

Table A132: MSD using estimated reference composite for the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.273	-0.182	-0.008	0.036
21-19	-0.932	0.006	-0.019	-0.086
25-15	-0.352	-0.700	-0.423	-0.305
40-0	-1.827	-1.721	-1.088	-0.794
0-40:40-0	-8.274	-6.283	-3.392	-2.658

Table A133: MSD using estimated reference composite for the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.168	0.118	0.128	0.045
21-19	0.750	0.047	0.177	0.060
25-15	1.668	1.186	0.651	0.396
40-0	5.588	3.967	2.089	1.753
0-40:40-0	8.863	6.568	4.216	3.043

Table A134: MSD using calculated reference composite for subpopulation 1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.083	-0.092	-0.004	0.065
21-19	-0.227	-0.053	0.037	-0.085
25-15	-0.044	-0.174	-0.078	-0.200
40-0	-0.109	-0.404	-0.640	-0.596
0-40:40-0	-1.292	-1.386	-1.332	-1.451

Table A135: MSD using calculated reference composite for subpopulation 2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.150	-0.073	0.028	0.076
21-19	-0.020	0.052	0.072	-0.039
25-15	0.565	0.326	0.382	0.312
40-0	1.260	1.073	0.941	0.947
0-40:40-0	2.012	1.912	2.016	1.882

Table A136: RMSD using estimated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.323	3.011	3.071	3.019
21-19	3.464	3.062	3.023	3.027
25-15	3.501	3.183	3.116	3.106
40-0	4.954	4.285	3.571	3.375
0-40:40-0	7.753	6.548	4.776	4.172

Table A137: RMSD using estimated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.041	1.830	1.873	1.813
21-19	2.087	1.813	1.770	1.849
25-15	1.948	1.782	1.732	1.830
40-0	1.981	1.840	1.800	1.795
0-40:40-0	2.191	2.346	1.897	1.950

Table A138: RMSD using estimated reference composite top 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.016	2.685	2.554	2.394
21-19	3.248	2.661	2.423	2.471
25-15	3.574	2.680	2.367	2.448
40-0	5.248	4.025	2.824	2.615
0-40:40-0	9.308	7.333	4.467	3.670

Table A139: RMSD using estimated reference composite top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.233	2.688	2.498	2.449
21-19	3.479	2.785	2.501	2.523
25-15	3.914	3.044	2.484	2.570
40-0	5.522	4.698	3.306	2.953
0-40:40-0	8.995	6.942	4.145	3.575

Table A140: RMSD using estimated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.388	2.321	2.394	2.452
21-19	2.367	2.308	2.510	2.393
25-15	2.250	2.414	2.589	2.367
40-0	2.756	2.631	2.535	2.576
0-40:40-0	3.162	3.264	2.974	2.827

Table A141: RMSD using estimated reference composite for the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.656	3.111	2.956	2.967
21-19	3.612	3.135	3.089	2.906
25-15	3.146	3.178	3.203	2.926
40-0	4.275	3.890	3.242	3.182
0-40:40-0	10.116	8.021	5.160	4.371

Table A142: RMSD using estimated reference composite for the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.202	2.930	2.943	3.050
21-19	3.452	2.991	3.136	2.932
25-15	3.868	3.462	3.298	3.022
40-0	7.137	5.593	3.952	3.720
0-40:40-0	10.561	8.392	5.692	4.544

Table A143: RMSD using calculated reference composite for subpopulation 1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.378	3.046	3.076	3.050
21-19	3.500	3.093	3.060	3.054
25-15	3.499	3.213	3.147	3.128
40-0	4.848	4.226	3.568	3.359
0-40:40-0	7.730	6.505	4.681	4.098

Table A144: RMSD using calculated reference composite for subpopulation 2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.266	2.975	3.066	2.986
21-19	3.426	3.029	2.985	2.998
25-15	3.501	3.151	3.083	3.083
40-0	5.056	4.341	3.573	3.388
0-40:40-0	7.775	6.587	4.865	4.235

Table A145: MSD using estimated reference composite full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.187	0.015	0.006	0.001
21-19	0.086	0.030	0.047	-0.012
25-15	0.378	0.190	0.055	0.097
40-0	0.971	0.679	0.481	0.327
0-40:40-0	0.718	0.630	0.740	0.711

Table A146: MSD using estimated reference composite top 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.034	0.225	-0.015	0.011
21-19	0.105	0.051	0.035	0.084
25-15	0.215	0.091	0.002	0.067
40-0	-0.330	-0.433	-0.239	-0.188
0-40:40-0	0.022	-0.069	-0.028	0.031

Table A147: MSD using estimated reference composite top 10% of $\theta 1$, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.511	0.102	-0.089	-0.048
21-19	0.285	0.153	0.170	0.205
25-15	1.582	1.316	0.559	0.471
40-0	3.579	2.307	1.524	1.218
0-40:40-0	7.524	5.863	3.419	2.654

Table A148: MSD using estimated reference composite top 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.449	0.145	0.085	-0.014
21-19	-0.202	-0.243	-0.187	-0.280
25-15	-1.436	-1.466	-0.729	-0.513
40-0	-3.825	-3.000	-2.244	-1.885
0-40:40-0	-6.970	-5.674	-3.490	-2.877

Table A149: MSD using estimated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.169	-0.242	-0.029	-0.017
21-19	-0.030	-0.053	0.021	-0.041
25-15	0.281	0.162	0.054	0.076
40-0	1.047	0.959	0.547	0.361
0-40:40-0	0.111	0.211	0.175	0.070

Table A150: MSD using estimated reference composite for the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.655	0.004	-0.067	-0.013
21-19	-0.048	-0.080	0.061	-0.032
25-15	-0.037	-0.139	-0.364	-0.227
40-0	-0.536	-1.129	-0.855	-0.757
0-40:40-0	-9.056	-6.899	-4.145	-3.086

Table A151: MSD using estimated reference composite for the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.177	-0.313	-0.020	0.156
21-19	0.158	0.237	0.060	-0.089
25-15	1.689	1.156	0.746	0.770
40-0	5.040	4.826	3.441	2.817
0-40:40-0	10.956	8.808	6.882	5.838

Table A152: MSD using calculated reference composite for subpopulation 1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.149	0.062	0.052	-0.021
21-19	0.051	-0.076	0.001	-0.087
25-15	-0.024	-0.274	-0.406	-0.320
40-0	-0.023	-0.479	-0.870	-1.049
0-40:40-0	-2.483	-2.583	-2.505	-2.460

Table A153: MSD using calculated reference composite for subpopulation 2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.225	-0.031	-0.040	0.023
21-19	0.121	0.135	0.092	0.063
25-15	0.780	0.654	0.516	0.513
40-0	1.965	1.838	1.831	1.703
0-40:40-0	3.918	3.843	3.985	3.883

Table A154: RMSD using estimated reference composite full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.385	3.097	3.002	2.986
21-19	3.208	3.053	3.014	3.024
25-15	3.318	3.185	3.097	3.077
40-0	4.656	4.291	3.787	3.634
0-40:40-0	8.426	7.046	5.716	5.119

Table A155: RMSD using estimated reference composite top 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.036	1.820	1.715	1.632
21-19	1.713	1.706	1.689	1.646
25-15	1.784	1.658	1.628	1.608
40-0	1.611	1.556	1.426	1.484
0-40:40-0	1.931	1.841	1.637	1.593

Table A156: RMSD using estimated reference composite top 10% of $\theta 1$, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.122	2.592	2.412	2.332
21-19	2.878	2.477	2.356	2.371
25-15	3.011	2.735	2.328	2.305
40-0	4.811	3.631	2.679	2.433
0-40:40-0	9.473	7.482	4.598	3.696

Table A157: RMSD using estimated reference composite top 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.787	2.882	2.523	2.453
21-19	3.571	2.912	2.566	2.482
25-15	3.749	3.299	2.682	2.559
40-0	5.419	4.352	3.408	3.108
0-40:40-0	9.023	7.147	4.551	3.762

Table A158: RMSD using estimated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.586	2.503	2.491	2.493
21-19	2.335	2.438	2.491	2.525
25-15	2.487	2.451	2.456	2.534
40-0	3.093	3.000	2.900	2.752
0-40:40-0	3.765	3.515	3.329	3.219

Table A159: RMSD using estimated reference composite for the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.465	3.312	3.153	3.114
21-19	3.261	3.157	3.121	3.125
25-15	3.181	3.088	3.175	3.145
40-0	3.890	3.897	3.596	3.412
0-40:40-0	11.123	8.766	6.065	5.007

Table A160: RMSD using estimated reference composite for the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.327	3.112	3.188	3.059
21-19	3.189	3.136	3.182	3.269
25-15	3.682	3.462	3.374	3.377
40-0	6.430	6.266	5.036	4.611
0-40:40-0	12.704	10.295	8.121	6.955

Table A161: RMSD using calculated reference composite for subpopulation 1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.505	3.241	3.104	3.100
21-19	3.337	3.190	3.133	3.131
25-15	3.407	3.281	3.205	3.169
40-0	4.550	4.188	3.722	3.626
0-40:40-0	8.324	6.823	5.256	4.603

Table A162: RMSD using calculated reference composite for subpopulation 2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.254	2.944	2.895	2.866
21-19	3.071	2.908	2.888	2.912
25-15	3.222	3.084	2.984	2.981
40-0	4.755	4.389	3.843	3.633
0-40:40-0	8.521	7.256	6.128	5.577

Table A163: MSD using estimated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.015	0.029	-0.023	-0.048
21-19	0.034	0.006	-0.005	-0.093
25-15	0.077	-0.013	0.047	-0.061
40-0	0.090	0.082	0.022	0.090
0-40:40-0	-0.060	0.105	0.047	0.069

Table A164: MSD using estimated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.050	-0.087	0.070	0.036
21-19	-0.034	-0.005	-0.025	-0.020
25-15	-0.153	0.061	-0.058	0.084
40-0	-0.688	-0.418	-0.182	-0.111
0-40:40-0	0.067	-0.105	0.048	0.007

Table A165: MSD using estimated reference composite top 10% of $\theta 1$, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.050	0.005	0.011	-0.090
21-19	0.176	0.032	0.026	-0.013
25-15	0.736	0.548	0.312	0.304
40-0	3.075	2.290	1.236	0.946
0-40:40-0	7.787	5.658	3.098	2.106

Table A166: MSD using estimated reference composite top 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.036	-0.023	0.029	0.059
21-19	-0.198	-0.080	-0.069	-0.024
25-15	-0.981	-0.593	-0.454	-0.196
40-0	-4.549	-3.271	-1.641	-1.070
0-40:40-0	-7.729	-5.527	-3.079	-1.986

Table A167: MSD using estimated reference composite for bottom 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.070	0.100	0.029	-0.012
21-19	0.174	0.023	-0.004	-0.092
25-15	0.213	0.141	0.221	-0.130
40-0	1.226	0.746	0.269	0.192
0-40:40-0	-0.057	-0.023	0.060	-0.055

Table A168: MSD using estimated reference composite for the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.116	0.057	-0.048	-0.132
21-19	-0.094	-0.162	-0.040	-0.187
25-15	-0.788	-0.733	-0.107	-0.370
40-0	-3.041	-2.351	-1.359	-0.941
0-40:40-0	-7.816	-5.943	-3.027	-2.123

Table A169: MSD using estimated reference composite for the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.092	0.084	0.040	0.151
21-19	0.287	0.222	0.074	0.030
25-15	1.245	0.841	0.558	0.076
40-0	4.946	3.553	1.904	1.283
0-40:40-0	7.553	6.050	3.158	2.240

Table A170: MSD using calculated reference composite for subpopulation 1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.010	0.030	-0.025	-0.055
21-19	0.044	0.003	0.001	-0.101
25-15	0.064	-0.030	0.042	-0.063
40-0	0.095	0.062	0.026	0.077
0-40:40-0	-0.064	0.089	0.032	0.080

Table A171: MSD using calculated reference composite for subpopulation 2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.020	0.028	-0.020	-0.040
21-19	0.025	0.010	-0.011	-0.085
25-15	0.090	0.003	0.051	-0.060
40-0	0.086	0.102	0.018	0.102
0-40:40-0	-0.056	0.121	0.061	0.059

Table A172: RMSD using estimated reference composite full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.266	3.263	3.382	3.350
21-19	3.272	3.318	3.380	3.336
25-15	3.409	3.357	3.413	3.376
40-0	4.989	4.424	3.772	3.633
0-40:40-0	7.538	6.338	4.586	4.084

Table A173: RMSD using estimated reference composite top 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	1.819	1.828	1.781	1.814
21-19	1.797	1.801	1.848	1.972
25-15	1.787	1.861	1.951	1.996
40-0	2.122	2.018	1.946	1.937
0-40:40-0	2.620	2.313	2.274	2.062

Table A174: RMSD using estimated reference composite top 10% of $\theta 1$, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.664	2.530	2.450	2.336
21-19	2.662	2.580	2.470	2.538
25-15	2.728	2.636	2.538	2.510
40-0	4.635	3.765	2.810	2.602
0-40:40-0	9.402	7.143	4.399	3.415

Table A175: RMSD using estimated reference composite top 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.649	2.613	2.410	2.405
21-19	2.685	2.578	2.510	2.548
25-15	2.905	2.745	2.637	2.583
40-0	5.886	4.702	3.188	2.833
0-40:40-0	9.395	6.989	4.374	3.346

Table A176: RMSD using estimated reference composite for bottom 10% of the full population, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.489	3.554	3.649	3.608
21-19	3.471	3.576	3.628	3.587
25-15	3.578	3.524	3.693	3.583
40-0	4.047	3.873	3.773	3.828
0-40:40-0	3.989	4.130	3.848	3.841

Table A177: RMSD using estimated reference composite for the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.546	3.535	3.783	3.738
21-19	3.542	3.643	3.779	3.670
25-15	3.737	3.712	3.796	3.708
40-0	5.586	4.865	4.212	3.980
0-40:40-0	9.726	7.860	5.080	4.500

Table A178: RMSD using estimated reference composite for the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.546	3.607	3.658	3.691
21-19	3.580	3.687	3.762	3.672
25-15	3.826	3.747	3.722	3.680
40-0	6.561	5.494	4.435	4.108
0-40:40-0	9.525	7.939	5.150	4.587

Table A179: RMSD using calculated reference composite for subpopulation 1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.270	3.246	3.366	3.361
21-19	3.276	3.309	3.379	3.342
25-15	3.411	3.355	3.411	3.373
40-0	4.995	4.419	3.761	3.646
0-40:40-0	7.533	6.332	4.578	4.086

Table A180: RMSD using calculated reference composite for subpopulation 2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.263	3.278	3.397	3.339
21-19	3.268	3.327	3.380	3.329
25-15	3.406	3.358	3.413	3.379
40-0	4.983	4.429	3.782	3.619
0-40:40-0	7.543	6.344	4.594	4.081

Table A181: MSD using estimated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.006	-0.002	0.088	-0.027
21-19	0.021	0.064	-0.043	-0.027
25-15	0.084	0.063	0.019	0.048
40-0	0.240	0.250	0.192	0.174
0-40:40-0	0.275	0.249	0.365	0.222

Table A182: MSD using estimated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.072	0.054	-0.093	0.004
21-19	0.027	-0.024	0.129	-0.006
25-15	-0.091	-0.057	-0.058	-0.091
40-0	-0.590	-0.465	-0.174	-0.129
0-40:40-0	-0.266	-0.161	-0.143	-0.086

Table A183: MSD using estimated reference composite top 10% of $\theta 1$, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.086	0.076	0.035	0.019
21-19	0.215	0.146	0.152	0.070
25-15	0.856	0.629	0.219	0.093
40-0	2.966	2.262	1.477	1.030
0-40:40-0	7.245	5.388	2.871	2.307

Table A184: MSD using estimated reference composite top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.032	0.015	-0.065	-0.019
21-19	-0.194	-0.136	-0.015	-0.132
25-15	-1.091	-0.823	-0.412	-0.364
40-0	-4.434	-3.295	-1.907	-1.289
0-40:40-0	-7.290	-5.572	-3.098	-2.455

Table A185: MSD using estimated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.056	-0.099	0.209	-0.057
21-19	0.055	0.110	-0.096	-0.090
25-15	0.198	0.375	-0.082	0.104
40-0	1.617	1.282	0.406	0.391
0-40:40-0	0.261	0.050	0.174	-0.100

Table A186: MSD using estimated reference composite for the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.081	-0.028	0.318	-0.097
21-19	-0.089	-0.039	-0.229	-0.169
25-15	-0.779	-0.426	-0.503	-0.240
40-0	-3.162	-2.215	-1.414	-0.928
0-40:40-0	-8.535	-6.380	-3.601	-2.668

Table A187: MSD using estimated reference composite for the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.076	-0.158	0.075	-0.065
21-19	0.237	0.196	0.046	0.040
25-15	1.177	1.038	0.616	0.591
40-0	5.669	4.474	2.429	1.918
0-40:40-0	9.502	7.109	4.677	3.387

Table A188: MSD using calculated reference composite for subpopulation 1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.011	-0.006	0.112	-0.047
21-19	-0. 01618128	0.041	-0.097	-0.067
25-15	-0.103	-0.122	-0.179	-0.140
40-0	-0.504	-0.496	-0.608	-0.645
0-40:40-0	-1.357	-1.432	-1.278	-1.445

Table A189: MSD using calculated reference composite for subpopulation 2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.001	0.002	0.064	-0.007
21-19	0.058	0.087	0.011	0.013
25-15	0.271	0.248	0.216	0.236
40-0	0.984	0.997	0.991	0.992
0-40:40-0	1.906	1.929	2.009	1.888

Table A190: RMSD using estimated reference composite full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.231	3.271	3.380	3.212
21-19	3.227	3.318	3.301	3.354
25-15	3.423	3.433	3.366	3.362
40-0	5.089	4.551	3.868	3.708
0-40:40-0	7.916	6.679	4.997	4.471

Table A191: RMSD using estimated reference composite top 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	1.710	1.711	1.721	1.635
21-19	1.701	1.704	1.798	1.801
25-15	1.651	1.692	1.705	1.769
40-0	1.848	1.788	1.731	1.670
0-40:40-0	2.104	2.058	1.839	1.944

Table A192: RMSD using estimated reference composite top 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.583	2.525	2.454	2.452
21-19	2.555	2.522	2.451	2.397
25-15	2.693	2.586	2.360	2.379
40-0	4.436	3.624	2.810	2.472
0-40:40-0	9.116	6.978	4.116	3.583

Table A193: RMSD using estimated reference composite top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.555	2.463	2.382	2.458
21-19	2.514	2.502	2.450	2.377
25-15	2.930	2.747	2.438	2.409
40-0	5.776	4.629	3.301	2.833
0-40:40-0	8.957	7.007	4.224	3.600

Table A194: RMSD using estimated reference composite for bottom 10% of the full population, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.564	3.648	3.746	3.668
21-19	3.511	3.673	3.615	3.816
25-15	3.598	3.717	3.778	3.778
40-0	4.390	4.243	4.023	3.922
0-40:40-0	4.574	4.517	4.344	4.171

Table A195: RMSD using estimated reference composite for the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.616	3.667	3.885	3.715
21-19	3.591	3.731	3.727	3.900
25-15	3.872	3.831	3.914	3.817
40-0	5.885	5.106	4.406	4.140
0-40:40-0	10.591	8.507	5.915	5.023

Table A196: RMSD using estimated reference composite for the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.585	3.690	3.920	3.642
21-19	3.553	3.704	3.624	3.894
25-15	3.844	3.913	3.953	3.887
40-0	7.191	6.223	4.791	4.457
0-40:40-0	11.308	8.964	6.593	5.494

Table A197: RMSD using calculated reference composite for subpopulation 1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.297	3.335	3.452	3.351
21-19	3.290	3.375	3.357	3.428
25-15	3.489	3.498	3.445	3.430
40-0	5.157	4.626	3.923	3.775
0-40:40-0	7.914	6.693	4.922	4.423

Table A198: RMSD using calculated reference composite for subpopulation 2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.163	3.204	3.307	3.066
21-19	3.162	3.259	3.243	3.278
25-15	3.355	3.367	3.284	3.293
40-0	5.018	4.474	3.812	3.637
0-40:40-0	7.917	6.661	5.066	4.512

Table A199: MSD using estimated reference composite full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.003	-0.027	0.049	0.030
21-19	0.054	-0.035	0.056	0.043
25-15	0.213	0.108	0.061	0.119
40-0	0.613	0.448	0.433	0.565
0-40:40-0	0.651	0.637	0.743	0.694

Table A200: MSD using estimated reference composite top 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.080	0.004	-0.002	0.050
21-19	-0.073	0.101	0.020	0.012
25-15	-0.101	-0.088	-0.068	-0.037
40-0	-0.585	-0.410	-0.247	-0.230
0-40:40-0	-0.296	-0.226	-0.211	-0.173

Table A201: MSD using estimated reference composite top 10% of $\theta 1$, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.063	0.019	0.047	0.073
21-19	0.031	0.257	0.029	0.023
25-15	0.804	0.635	0.318	0.359
40-0	3.004	2.384	1.396	1.083
0-40:40-0	7.394	5.383	3.150	2.381

Table A202: MSD using estimated reference composite top 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.024	-0.019	-0.015	0.006
21-19	-0.086	-0.161	-0.077	0.041
25-15	-1.006	-0.912	-0.526	-0.455
40-0	-4.890	-3.606	-2.316	-2.016
0-40:40-0	-7.502	-5.602	-3.736	-2.983

Table A203: MSD using estimated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.002	-0.057	0.065	-0.014
21-19	0.124	-0.150	0.173	0.128
25-15	0.521	0.373	0.003	0.193
40-0	3.205	1.565	1.147	1.238
0-40:40-0	0.897	0.736	0.829	0.630

Table A204: MSD using estimated reference composite for the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.074	-0.097	0.010	0.039
21-19	-0.065	-0.285	0.023	-0.043
25-15	-0.672	-0.496	-0.459	-0.172
40-0	-2.358	-2.209	-1.195	-0.443
0-40:40-0	-9.262	-7.278	-4.063	-3.165

Table A205: MSD using estimated reference composite for the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.006	-0.065	0.096	-0.049
21-19	0.416	0.134	0.362	0.238
25-15	1.629	1.200	0.810	0.825
40-0	7.121	5.291	4.047	3.908
0-40:40-0	11.739	10.105	7.849	6.897

Table A206: MSD using calculated reference composite for subpopulation 1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	-0.003	-0.047	0.015	0.000
21-19	0.025	-0.146	-0.007	-0.028
25-15	-0.106	-0.242	-0.316	-0.243
40-0	-0.671	-0.965	-1.021	-0.903
0-40:40-0	-2.634	-2.629	-2.496	-2.530

Table A207: MSD using calculated reference composite for subpopulation 2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	0.009	-0.007	0.084	0.059
21-19	0.083	0.076	0.119	0.115
25-15	0.531	0.457	0.439	0.480
40-0	1.896	1.862	1.887	2.033
0-40:40-0	3.936	3.903	3.981	3.918

Table A208: RMSD using estimated reference composite full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.194	3.212	3.310	3.362
21-19	3.182	3.221	3.326	3.379
25-15	3.462	3.382	3.343	3.389
40-0	5.469	4.978	4.168	4.133
0-40:40-0	8.702	7.458	6.081	5.551

Table A209: RMSD using estimated reference composite top 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	1.659	1.635	1.624	1.686
21-19	1.595	1.613	1.637	1.333
25-15	1.552	1.603	1.551	1.581
40-0	1.569	4.770	1.418	1.335
0-40:40-0	1.762	1.650	1.579	1.495

Table A210: RMSD using estimated reference composite top 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.540	2.452	2.410	2.414
21-19	2.505	2.480	2.397	2.179
25-15	2.654	2.557	2.322	2.341
40-0	4.529	1.495	2.602	2.275
0-40:40-0	9.519	7.240	4.500	3.526

Table A211: RMSD using estimated reference composite top 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	2.529	2.458	2.351	2.403
21-19	2.505	2.451	2.388	1.927
25-15	2.872	2.719	2.383	2.367
40-0	6.206	3.672	3.487	3.106
0-40:40-0	9.294	7.071	4.757	3.848

Table A212: RMSD using estimated reference composite for bottom 10% of the full population, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.512	3.668	3.801	3.812
21-19	3.610	3.667	3.824	3.919
25-15	3.778	3.818	3.908	3.859
40-0	5.399	4.890	4.408	4.499
0-40:40-0	5.135	5.013	4.916	4.871

Table A213: RMSD using estimated reference composite for the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.629	3.715	3.885	3.988
21-19	3.619	3.706	3.859	3.978
25-15	3.960	3.879	3.952	3.972
40-0	5.990	4.618	4.590	4.584
0-40:40-0	11.613	9.588	6.702	5.807

Table A214: RMSD using estimated reference composite for the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.497	3.642	3.800	3.915
21-19	3.542	3.646	3.837	3.883
25-15	4.070	3.877	3.935	4.014
40-0	8.458	5.461	5.882	5.870
0-40:40-0	13.459	11.766	9.408	8.389

Table A215: RMSD using estimated reference composite for subpopulation 1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.327	3.351	3.447	3.516
21-19	3.316	3.355	3.473	3.528
25-15	3.612	3.526	3.503	3.540
40-0	5.587	6.931	4.236	4.142
0-40:40-0	8.650	7.296	5.701	5.126

Table A216: RMSD using estimated reference composite for subpopulation 2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	3.053	3.066	3.165	3.201
21-19	3.042	3.080	3.171	3.222
25-15	3.303	3.229	3.174	3.230
40-0	5.347	4.881	4.094	4.118
0-40:40-0	8.751	7.611	6.430	5.936

APPENDIX B

Output of proc glm

In the output below “calc_est” is an indicator variable whether the error was obtained using the calculated or the estimated reference composite, “split” indicates the difference between the forms – 0, 1, 5, 10 and 40 items, “difference” indicates the difference between subpopulation means.

3PL results are presented first, followed by the 1PL results on page 178.

3pl

---------- stat=10 ----------

The GLM Procedure
Class Level Information

Class	Levels	Values
calc_est	2	1 2
correlation	4	0.5 0.7 0.9 0.95
split	5	0 1 5 20 40
difference	3	0 0.5 1

Number of Observations Read	108
Number of Observations Used	108

Dependent Variable: msd msd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	8.63349166	0.14389153	3.41	<.0001
Error	47	1.98112681	0.04215163		
Corrected Total	107	10.61461847			

R-Square	Coeff Var	Root MSE	msd Mean
0.813359	-293.3956	0.205309	-0.069977

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	0.32915894	0.32915894	7.81	0.0075
correlation	3	0.13291595	0.04430532	1.05	0.3789
split	4	6.05249710	1.51312427	35.90	<.0001
correlation*split	12	0.62345996	0.05195500	1.23	0.2903
difference	2	0.44591615	0.22295808	5.29	0.0085
correlati*difference	6	0.04879896	0.00813316	0.19	0.9773
split*difference	8	0.67434469	0.08429309	2.00	0.0671
correl*split*differe	24	0.32639991	0.01360000	0.32	0.9981

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	0.15707477	0.15707477	3.73	0.0596
correlation	3	0.09146296	0.03048765	0.72	0.5432
split	4	6.05249710	1.51312427	35.90	<.0001
correlation*split	12	0.62345996	0.05195500	1.23	0.2903
difference	2	0.28354930	0.14177465	3.36	0.0431
correlati*difference	6	0.09048213	0.01508035	0.36	0.9017
split*difference	8	0.67434469	0.08429309	2.00	0.0671
correl*split*differe	24	0.32639991	0.01360000	0.32	0.9981

Dependent Variable: rmsd rmsd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	67.60535218	1.12675587	9.37	<.0001
Error	47	5.65078674	0.12022951		
Corrected Total	107	73.25613892			

R-Square	Coeff Var	Root MSE	rmsd Mean
0.922863	26.55585	0.346741	1.305706

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	53.47702097	53.47702097	444.79	<.0001
correlation	3	1.25960706	0.41986902	3.49	0.0227
split	4	5.90116394	1.47529098	12.27	<.0001
correlation*split	12	0.33225456	0.02768788	0.23	0.9958
difference	2	4.76238605	2.38119303	19.81	<.0001
correlati*difference	6	0.27814944	0.04635824	0.39	0.8846
split*difference	8	1.21791888	0.15223986	1.27	0.2838
correl*split*differe	24	0.37685127	0.01570214	0.13	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	48.06512560	48.06512560	399.78	<.0001
correlation	3	1.14330128	0.38110043	3.17	0.0328
split	4	5.90116394	1.47529098	12.27	<.0001
correlation*split	12	0.33225456	0.02768788	0.23	0.9958
difference	2	4.10090332	2.05045166	17.05	<.0001

correlati*difference	6	0.20021725	0.03336954	0.28	0.9447
split*difference	8	1.21791888	0.15223986	1.27	0.2838
correl*split*differe	24	0.37685127	0.01570214	0.13	1.0000

Least Squares Means

calc_est	msd LSMEAN	rmsd LSMEAN
1	-0.10149827	0.51991874
2	-0.02059845	1.93509137

correlation	msd LSMEAN	rmsd LSMEAN
0.5	-0.07170901	1.39776573
0.7	-0.10574975	1.22872151
0.9	-0.03602982	1.17370133
0.95	-0.03070487	1.10983164

split	msd LSMEAN	rmsd LSMEAN
0	0.05975684	1.23128385
1	0.02130982	0.99278883
5	0.00879431	0.98604949
20	-0.51202527	1.36146063
40	0.11692250	1.56594247

difference	msd LSMEAN	rmsd LSMEAN
0	-0.13204012	1.48592748
0.5	-0.04730949	1.20519068
1	-0.00379548	0.99139699

---------- stat=90 ----------

The GLM Procedure
Class Level Information

Class	Levels	Values
calc_est	2	1 2
correlation	4	0.5 0.7 0.9 0.95
split	5	0 1 5 20 40
difference	3	0 0.5 1

Number of Observations Read 108
Number of Observations Used 108

Dependent Variable: msd msd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	28.81673300	0.48027888	4.29	<.0001
Error	47	5.26762881	0.11207721		
Corrected Total	107	34.08436181			

R-Square	Coeff Var	Root MSE	msd Mean
0.845453	117.8403	0.334779	0.284096

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	5.57408789	5.57408789	49.73	<.0001
correlation	3	0.16671905	0.05557302	0.50	0.6869
split	4	18.75161742	4.68790436	41.83	<.0001
correlation*split	12	1.44465191	0.12038766	1.07	0.4025
difference	2	1.49571131	0.74785566	6.67	0.0028
correlati*difference	6	0.02940833	0.00490139	0.04	0.9996
split*difference	8	1.07025444	0.13378181	1.19	0.3232
correl*split*differe	24	0.28428265	0.01184511	0.11	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F

calc_est	1	4.04904152	4.04904152	36.13	<.0001
correlation	3	0.11021488	0.03673829	0.33	0.8052
split	4	18.75161742	4.68790436	41.83	<.0001
correlation*split	12	1.44465191	0.12038766	1.07	0.4025
difference	2	1.16868551	0.58434275	5.21	0.0090
correlati*difference	6	0.05840321	0.00973387	0.09	0.9973
split*difference	8	1.07025444	0.13378181	1.19	0.3232
correl*split*differe	24	0.28428265	0.01184511	0.11	1.0000

Dependent Variable: rmsd rmsd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	104.9522386	1.7492040	5.00	<.0001
Error	47	16.4396826	0.3497805		
Corrected Total	107	121.3919212			

R-Square	Coeff Var	Root MSE	rmsd Mean
0.864574	30.22273	0.591422	1.956879

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	53.95686805	53.95686805	154.26	<.0001
correlation	3	0.83414794	0.27804931	0.79	0.5029
split	4	37.71071987	9.42767997	26.95	<.0001
correlation*split	12	1.80667659	0.15055638	0.43	0.9430
difference	2	5.50581825	2.75290913	7.87	0.0011
correlati*difference	6	0.01732285	0.00288714	0.01	1.0000
split*difference	8	4.80550590	0.60068824	1.72	0.1192
correl*split*differe	24	0.31517912	0.01313246	0.04	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	50.32892428	50.32892428	143.89	<.0001
correlation	3	0.57373600	0.19124533	0.55	0.6528
split	4	37.71071987	9.42767997	26.95	<.0001
correlation*split	12	1.80667659	0.15055638	0.43	0.9430
difference	2	4.09137505	2.04568752	5.85	0.0054
correlati*difference	6	0.04683585	0.00780598	0.02	0.9999

split*difference	8	4.80550590	0.60068824	1.72	0.1192
correl*split*differe	24	0.31517912	0.01313246	0.04	1.0000

Least Squares Means

calc_est	msd LSMEAN	rmsd LSMEAN
1	0.49164074	1.14096676
2	0.08089744	2.58908227

correlation	msd LSMEAN	rmsd LSMEAN
0.5	0.32419094	1.96402680
0.7	0.31435698	1.90393734
0.9	0.25204872	1.82865140
0.95	0.25447971	1.76348252

split	msd LSMEAN	rmsd LSMEAN
0	0.10045782	1.76238858
1	-0.02261733	1.20945686
5	0.09445921	1.34422045
20	1.07702820	2.34897517
40	0.18201754	2.66008152

difference	msd LSMEAN	rmsd LSMEAN
0	0.14522749	1.62994103
0.5	0.30569455	1.84141938
1	0.40788523	2.12371314

---------- stat=100 ----------

The GLM Procedure
Class Level Information

Class	Levels	Values
calc_est	2	1 2
correlation	4	0.5 0.7 0.9 0.95
split	5	0 1 5 20 40
difference	3	0 0.5 1

Number of Observations Read 108
Number of Observations Used 108

Dependent Variable: msd msd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	43.42327604	0.72372127	2.15	0.0036
Error	47	15.79527728	0.33606973		
Corrected Total	107	59.21855332			

R-Square	Coeff Var	Root MSE	msd Mean
0.733271	148.0152	0.579715	0.391659

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	7.31688375	7.31688375	21.77	<.0001
correlation	3	0.04397084	0.01465695	0.04	0.9877
split	4	14.97007035	3.74251759	11.14	<.0001
correlation*split	12	0.10293396	0.00857783	0.03	1.0000
difference	2	9.81151677	4.90575839	14.60	<.0001
correlati*difference	6	0.01214643	0.00202441	0.01	1.0000
split*difference	8	11.07527175	1.38440897	4.12	0.0009
correl*split*differe	24	0.09048219	0.00377009	0.01	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F

calc_est	1	5.68079711	5.68079711	16.90	0.0002
correlation	3	0.02620448	0.00873483	0.03	0.9943
split	4	14.97007035	3.74251759	11.14	<.0001
correlation*split	12	0.10293396	0.00857783	0.03	1.0000
difference	2	7.17965314	3.58982657	10.68	0.0001
correlati*difference	6	0.00965892	0.00160982	0.00	1.0000
split*difference	8	11.07527175	1.38440897	4.12	0.0009
correl*split*differe	24	0.09048219	0.00377009	0.01	1.0000

Dependent Variable: rmsd rmsd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	417.8426029	6.9640434	9.60	<.0001
Error	47	34.0889612	0.7252970		
Corrected Total	107	451.9315641			

R-Square	Coeff Var	Root MSE	rmsd Mean
0.924571	26.20733	0.851644	3.249640

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	56.7391842	56.7391842	78.23	<.0001
correlation	3	33.8134531	11.2711510	15.54	<.0001
split	4	277.1343836	69.2835959	95.52	<.0001
correlation*split	12	34.5915431	2.8826286	3.97	0.0003
difference	2	4.8396773	2.4198386	3.34	0.0442
correlati*difference	6	0.5240045	0.0873341	0.12	0.9934
split*difference	8	9.4840148	1.1855019	1.63	0.1406
correl*split*differe	24	0.7163423	0.0298476	0.04	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	64.8984686	64.8984686	89.48	<.0001
correlation	3	26.3468511	8.7822837	12.11	<.0001
split	4	277.1343836	69.2835959	95.52	<.0001
correlation*split	12	34.5915431	2.8826286	3.97	0.0003
difference	2	3.4703552	1.7351776	2.39	0.1025
correlati*difference	6	0.3531857	0.0588643	0.08	0.9978

split*difference	8	9.4840148	1.1855019	1.63	0.1406
correl*split*differe	24	0.7163423	0.0298476	0.04	1.0000

Least Squares Means

calc_est	msd LSMEAN	rmsd LSMEAN
1	0.64537024	2.25352186
2	0.15885212	3.89793752

correlation	msd LSMEAN	rmsd LSMEAN
0.5	0.42978574	3.80989927
0.7	0.39174259	3.28779767
0.9	0.39699668	2.70799941
0.95	0.38991971	2.49722239

split	msd LSMEAN	rmsd LSMEAN
0	0.25291989	2.33274875
1	0.02282473	1.63891724
5	0.16138595	1.98532073
20	0.54405786	3.54041963
40	1.02936748	5.88124208

difference	msd LSMEAN	rmsd LSMEAN
0	0.06826198	2.88186496
0.5	0.41378331	3.01817037
1	0.72428825	3.32715373

1pl

---------- stat=10 ----------

The GLM Procedure
Class Level Information

Class	Levels	Values
calc_est	2	1 2
correlation	4	0.5 0.7 0.9 0.95
split	5	0 1 5 20 40
difference	3	0 0.5 1

Number of Observations Read	108
Number of Observations Used	108

The GLM Procedure

Dependent Variable: msd msd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	7.39385388	0.12323090	1.98	0.0081
Error	47	2.92146220	0.06215877		
Corrected Total	107	10.31531608			

R-Square	Coeff Var	Root MSE	msd Mean
0.716784	-221.8111	0.249317	-0.112400

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	0.00964631	0.00964631	0.16	0.6954
correlation	3	0.37290136	0.12430045	2.00	0.1269
split	4	5.30417921	1.32604480	21.33	<.0001
correlation*split	12	0.59736970	0.04978081	0.80	0.6476
difference	2	0.30047305	0.15023653	2.42	0.1002
correlati*difference	6	0.02674780	0.00445797	0.07	0.9984
split*difference	8	0.61873021	0.07734128	1.24	0.2953
correl*split*differe	24	0.16380624	0.00682526	0.11	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	0.00189797	0.00189797	0.03	0.8620
correlation	3	0.28114883	0.09371628	1.51	0.2248
split	4	5.30417921	1.32604480	21.33	<.0001
correlation*split	12	0.59736970	0.04978081	0.80	0.6476
difference	2	0.22698388	0.11349194	1.83	0.1723
correlati*difference	6	0.02458881	0.00409813	0.07	0.9988
split*difference	8	0.61873021	0.07734128	1.24	0.2953
correl*split*differe	24	0.16380624	0.00682526	0.11	1.0000

Dependent Variable: rmsd rmsd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	97.5157321	1.6252622	0.95	0.5837
Error	47	80.7365982	1.7178000		
Corrected Total	107	178.2523303			

R-Square	Coeff Var	Root MSE	rmsd Mean
0.547066	86.96484	1.310649	1.507102

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	14.03299141	14.03299141	8.17	0.0063
correlation	3	4.33590032	1.44530011	0.84	0.4781
split	4	35.20377963	8.80094491	5.12	0.0016
correlation*split	12	6.19247814	0.51603985	0.30	0.9863
difference	2	8.23386402	4.11693201	2.40	0.1021
correlati*difference	6	3.56478258	0.59413043	0.35	0.9088
split*difference	8	19.65117298	2.45639662	1.43	0.2092
correl*split*differe	24	6.30076299	0.26253179	0.15	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	13.61432706	13.61432706	7.93	0.0071
correlation	3	3.28198517	1.09399506	0.64	0.5950
split	4	35.20377963	8.80094491	5.12	0.0016

correlation*split	12	6.19247814	0.51603985	0.30	0.9863
difference	2	6.04220485	3.02110243	1.76	0.1834
correlati*difference	6	2.63578811	0.43929802	0.26	0.9545
split*difference	8	19.65117298	2.45639662	1.43	0.2092
correl*split*differe	24	6.30076299	0.26253179	0.15	1.0000

Least Squares Means

calc_est	msd LSMEAN	rmsd LSMEAN
1	-0.09505459	1.07634264
2	-0.10394739	1.82951170

correlation	msd LSMEAN	rmsd LSMEAN
0.5	-0.17270182	1.66022535
0.7	-0.12556098	1.60194054
0.9	-0.06174726	1.31098554
0.95	-0.03799388	1.23855725

split	msd LSMEAN	rmsd LSMEAN
0	0.01214773	1.34193922
1	0.00580722	0.89550280
5	-0.04583723	1.02585470
20	-0.52268763	1.57110673
40	0.05306497	2.43023242

difference	msd LSMEAN	rmsd LSMEAN
0	-0.16246181	1.40586177
0.5	-0.08879820	1.17817862
1	-0.04724295	1.77474113

---------- stat=90 ----------

The GLM Procedure
Class Level Information

Class	Levels	Values
calc_est	2	1 2
correlation	4	0.5 0.7 0.9 0.95
split	5	0 1 5 20 40
difference	3	0 0.5 1

Number of Observations Read 108
Number of Observations Used 108

The GLM Procedure

Dependent Variable: msd msd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	79.03325705	1.31722095	4.90	<.0001
Error	47	12.63979115	0.26893173		
Corrected Total	107	91.67304820			

R-Square	Coeff Var	Root MSE	msd Mean
0.862121	89.53374	0.518586	0.579208

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	8.93287894	8.93287894	33.22	<.0001
correlation	3	1.19142163	0.39714054	1.48	0.2330
split	4	55.50640482	13.87660121	51.60	<.0001
correlation*split	12	4.31855827	0.35987986	1.34	0.2299
difference	2	4.32456772	2.16228386	8.04	0.0010
correlati*difference	6	0.12020635	0.02003439	0.07	0.9983
split*difference	8	4.08791389	0.51098924	1.90	0.0824
correl*split*differe	24	0.55130542	0.02297106	0.09	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	6.03302052	6.03302052	22.43	<.0001
correlation	3	0.87201644	0.29067215	1.08	0.3664
split	4	55.50640482	13.87660121	51.60	<.0001
correlation*split	12	4.31855827	0.35987986	1.34	0.2299
difference	2	3.22152740	1.61076370	5.99	0.0048
correlati*difference	6	0.09444122	0.01574020	0.06	0.9991
split*difference	8	4.08791389	0.51098924	1.90	0.0824
correl*split*differe	24	0.55130542	0.02297106	0.09	1.0000

Dependent Variable: rmsd rmsd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	228.7437353	3.8123956	3.93	<.0001
Error	47	45.5875171	0.9699472		
Corrected Total	107	274.3312524			

R-Square	Coeff Var	Root MSE	rmsd Mean
0.833823	32.23768	0.984859	3.054993

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	107.6265122	107.6265122	110.96	<.0001
correlation	3	2.3210507	0.7736836	0.80	0.5014
split	4	96.4767987	24.1191997	24.87	<.0001
correlation*split	12	4.1022289	0.3418524	0.35	0.9734
difference	2	10.1933756	5.0966878	5.25	0.0087
correlati*difference	6	0.0341267	0.0056878	0.01	1.0000
split*difference	8	7.6968889	0.9621111	0.99	0.4546
correl*split*differe	24	0.2927535	0.0121981	0.01	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	104.3450687	104.3450687	107.58	<.0001
correlation	3	1.5552841	0.5184280	0.53	0.6609
split	4	96.4767987	24.1191997	24.87	<.0001
correlation*split	12	4.1022289	0.3418524	0.35	0.9734

difference	2	7.9214455	3.9607228	4.08	0.0232
correlati*difference	6	0.0231972	0.0038662	0.00	1.0000
split*difference	8	7.6968889	0.9621111	0.99	0.4546
correl*split*differe	24	0.2927535	0.0121981	0.01	1.0000

Least Squares Means

calc_est	msd LSMEAN	rmsd LSMEAN
1	0.82334734	1.86275568
2	0.32197338	3.94787227

correlation	msd LSMEAN	rmsd LSMEAN
0.5	0.71761386	3.05493783
0.7	0.58619655	2.99703213
0.9	0.51714757	2.81847240
0.95	0.46968345	2.75081353

split	msd LSMEAN	rmsd LSMEAN
0	0.26304773	2.60076120
1	0.02561301	1.87418386
5	0.21039447	2.11665979
20	1.92489912	3.82004475
40	0.43934747	4.11492025

difference	msd LSMEAN	rmsd LSMEAN
0	0.35887720	2.57193197
0.5	0.56103831	2.88368606
1	0.79806557	3.26032389

---------- stat=100 ----------

The GLM Procedure
Class Level Information

Class	Levels	Values
calc_est	2	1 2
correlation	4	0.5 0.7 0.9 0.95
split	5	0 1 5 20 40
difference	3	0 0.5 1

Number of Observations Read	108
Number of Observations Used	108

Dependent Variable: msd msd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	47.97243076	0.79954051	2.15	0.0036
Error	47	17.44313544	0.37113054		
Corrected Total	107	65.41556620			

R-Square	Coeff Var	Root MSE	msd Mean
0.733349	149.7970	0.609205	0.406687

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	9.93160785	9.93160785	26.76	<.0001
correlation	3	0.01032026	0.00344009	0.01	0.9988
split	4	16.33114782	4.08278696	11.00	<.0001
correlation*split	12	0.05474757	0.00456230	0.01	1.0000
difference	2	10.34886469	5.17443234	13.94	<.0001
correlati*difference	6	0.01940768	0.00323461	0.01	1.0000
split*difference	8	11.24385350	1.40548169	3.79	0.0017
correl*split*differe	24	0.03248140	0.00135339	0.00	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F

calc_est	1	8.02885076	8.02885076	21.63	<.0001
correlation	3	0.00995038	0.00331679	0.01	0.9988
split	4	16.33114782	4.08278696	11.00	<.0001
correlation*split	12	0.05474757	0.00456230	0.01	1.0000
difference	2	7.80863818	3.90431909	10.52	0.0002
correlati*difference	6	0.02119833	0.00353305	0.01	1.0000
split*difference	8	11.24385350	1.40548169	3.79	0.0017
correl*split*differe	24	0.03248140	0.00135339	0.00	1.0000

The GLM Procedure

Dependent Variable: rmsd rmsd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	60	421.7917484	7.0298625	4.79	<.0001
Error	47	69.0483800	1.4691145		
Corrected Total	107	490.8401284			

R-Square	Coeff Var	Root MSE	rmsd Mean
0.859326	40.17844	1.212070	3.016718

Source	DF	Type I SS	Mean Square	F Value	Pr > F
calc_est	1	163.0553838	163.0553838	110.99	<.0001
correlation	3	20.9536359	6.9845453	4.75	0.0056
split	4	174.7907752	43.6976938	29.74	<.0001
correlation*split	12	29.3579349	2.4464946	1.67	0.1060
difference	2	14.2341336	7.1170668	4.84	0.0122
correlati*difference	6	3.1669280	0.5278213	0.36	0.9008
split*difference	8	12.6316633	1.5789579	1.07	0.3968
correl*split*differe	24	3.6012935	0.1500539	0.10	1.0000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
calc_est	1	172.4011234	172.4011234	117.35	<.0001
correlation	3	15.3751663	5.1250554	3.49	0.0228
split	4	174.7907752	43.6976938	29.74	<.0001
correlation*split	12	29.3579349	2.4464946	1.67	0.1060

difference	2	10.7253992	5.3626996	3.65	0.0336
correlati*difference	6	2.4274618	0.4045770	0.28	0.9457
split*difference	8	12.6316633	1.5789579	1.07	0.3968
correl*split*differe	24	3.6012935	0.1500539	0.10	1.0000

Least Squares Means

calc_est	msd LSMEAN	rmsd LSMEAN
1	0.71384421	1.43554303
2	0.13545381	4.11572588

correlation	msd LSMEAN	rmsd LSMEAN
0.5	0.41912133	3.31931227
0.7	0.41326386	2.96558462
0.9	0.43999079	2.49318582
0.95	0.42622007	2.32445511

split	msd LSMEAN	rmsd LSMEAN
0	0.29711151	1.94597124
1	0.02763929	1.70414234
5	0.15206581	1.95929272
20	0.57347633	3.26320082
40	1.07295212	5.00556516

difference	msd LSMEAN	rmsd LSMEAN
0	0.08112334	2.91112444
0.5	0.42723100	3.09144237
1	0.76559269	2.32433655

APPENDIX C

Exploratory tables of the number of examinees at the extremes of the population for the last simulation run.

Table C1: Number of examinees in the bottom 10% of both subpopulations, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	135	46	40	68
21-19	129	207	280	330
25-15	141	192	283	307
40-0	124	196	265	300
0-40:40-0	123	191	270	280

Table C2: Number of examinees in the bottom 10% of θ1, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	271	45	22	19
21-19	268	20	139	98
25-15	253	231	116	84
40-0	256	188	119	98
0-40:40-0	247	233	118	88

Table C3: Number of examinees in the bottom 10% of θ2, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	272	61	57	20
21-19	283	211	114	82
25-15	258	231	143	117
40-0	272	233	120	91
0-40:40-0	262	219	113	83

Table C4: Number of examinees in the top 10% of both subpopulations, no difference in subpopulation means

Top 10%				
item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	118	42	85	80
21-19	113	189	276	334
25-15	135	183	244	289
40-0	136	178	282	322
0-40:40-0	117	168	269	334

Table C5: Number of examinees in the top 10% of $\theta 1$, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	262	49	86	21
21-19	254	213	129	85
25-15	253	219	141	78
40-0	254	209	93	85
0-40:40-0	267	183	113	94

Table C6: Number of examinees in the top 10% of $\theta 2$, no difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	260	66	21	27
21-19	288	205	116	87
25-15	304	226	138	97
40-0	271	207	121	86
0-40:40-0	271	215	135	96

Table C7: Number of examinees in the bottom 10% of both subpopulations, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	128	176	328	300
21-19	137	127	242	321
25-15	122	142	271	291
40-0	153	146	271	305
0-40:40-0	133	146	283	309

Table C8: Number of examinees in the bottom 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	276	223	267	117
21-19	294	129	167	119
25-15	295	159	140	142
40-0	342	110	160	128
0-40:40-0	290	110	166	125

Table C9: Number of examinees in the bottom 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	267	216	63	94
21-19	288	262	121	104
25-15	240	267	140	116
40-0	266	254	123	104
0-40:40-0	254	254	133	112

Table C10: Number of examinees in the top 10% of both subpopulations, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	135	184	217	310
21-19	136	224	284	286
25-15	124	264	259	307
40-0	147	251	278	323
0-40:40-0	120	254	266	300

Table C11: Number of examinees in the top 10% of θ1, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	320	223	49	110
21-19	285	299	152	125
25-15	314	390	163	133
40-0	297	408	158	122
0-40:40-0	325	408	129	130

Table C12: Number of examinees in the top 10% of θ2, 0.5 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	266	187	202	88
21-19	274	163	124	111
25-15	280	152	107	91
40-0	289	167	129	105
0-40:40-0	277	167	122	120

Table C13: Number of examinees in the bottom 10% of both subpopulations, 1 difference in subpopulation means

Bottom 10%				
item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	139	193	226	251
21-19	77	194	255	262
25-15	78	197	141	297
40-0	66	197	225	244
0-40:40-0	66	197	225	294

Table C14: Number of examinees in the bottom 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	416	325	226	234
21-19	135	318	271	245
25-15	161	328	151	248
40-0	155	306	254	251
0-40:40-0	151	300	250	246

Table C15: Number of examinees in the bottom 10% of θ2, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	282	240	140	127
21-19	352	200	145	133
25-15	302	203	139	118
40-0	291	185	128	131
0-40:40-0	322	204	140	113

Table C16: Number of examinees in the top 10% of both subpopulations, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	136	172	267	260
21-19	206	199	252	284
25-15	212	195	272	265
40-0	222	187	263	257
0-40:40-0	209	89	237	248

Table C17: Number of examinees in the top 10% of θ1, 1 difference in subpopulation means

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	388	314	236	249
21-19	760	336	284	252
25-15	754	319	243	250
40-0	767	305	256	266
0-40:40-0	740	319	266	260

Table C18: Number of examinees in the top 10% of θ2, 1 difference in subpopulation means

Top 10% θ_2

item distribution	Correlation			
	0.5	0.7	0.9	0.95
20-20	243	212	144	139
21-19	181	180	138	127
25-15	179	193	152	132
40-0	200	235	154	128
0-40:40-0	174	219	139	132

Appendix D

Parscale code used for estimation

Examinee parameter estimation in the 3PL model:

```
>FILE DFNAME='filename',SAVE;
>SAVE PARM='rsamp.par',SCORE='rsamp.sco';
>INPUT NIDCHAR=6,NTOTAL=40;
(6A1,40A1)
>TEST    NBLOCK=1;
>BLOCK1 NITEMS=40,NCAT=2,ORIGINAL=(0,1),GUESSING=(2,ESTIMATE);
>CAL PARTIAL,LOGISTIC, DIST=2, NQPT=40, CRIT=(0.001),NEWTON=5, SPRIOR,
TPRIOR, GPRIOR;
>SCORE   EAP, ITERATION=(0.001, 40),PRINT;
```

For the 1PL (Rasch) model, the following code was used:

```
>FILE DFNAME='c:/pdiss/rsamp.dat',SAVE;
>SAVE PARM='c:/pdiss/rsamp.par',SCORE='c:/pdiss/rsamp.sco';
>INPUT NIDCHAR=6,NTOTAL=40;
(6A1,40A1)
>TEST    NBLOCK=1;
>BLOCK1 NITEMS=40,NCAT=2,ORIGINAL=(0,1),GUESSING=(2,FIX);
>CAL GRADED,LOGISTIC,SPRIOR,GPRIOR, CSLOPE;
>SCORE   EAP,PRINT;
```

Appendix E
Curriculum Vita

DOROTA STANIEWSKA

EDUCATION
1996 – 2000 Smith College, Northampton, MA
BA degree, Mathematics, minor in Medieval History

University of Hamburg, Hamburg, Germany
1998/1999 Junior Year Abroad program from Smith College

2000-2002 University of South Carolina, Columbia, SC
MS degree, Statistics,

Fall 2003 – Spring 2009, Rutgers University, New Brunswick, NJ
Ph.D program in Educational Measurement (Graduate School of Education)

WORK AND RESEARCH EXPERIENCE
Statistics Laboratory Assistant, USC, Columbia, SC (Fall 2000)

Teaching Assistant, USC, Columbia, SC (Spring 2001)

Research/Programming Assistant, USC, Columbia, SC (Summer 2001)

Research Assistant, USC, Columbia, SC (Fall 2001-Spring 2002)

Health Program Analyst, UMDNJ SPH, New Brunswick, NJ (July 2002- August 2004)

Research Associate, Educational Testing Service, Princeton, NJ (August 2004 – March 2007)

Senior Research Associate, Educational Testing Service, Princeton, NJ (March 2007 – present)

PUBLICATIONS

Hrywna M, Delnevo CD, Staniewska D. *Prevalence and correlates of Internet cigarette purchasing among adult smokers in New Jersey*. Tobacco Control. 2004;13(3):296-300.