Running Head: FIT STATISTICS IN SMALL SAMPLE PILOT TESTS

Using Rasch Fit Statistics in Small Sample Pilot Tests:

Promise or Peril?

Adam VanIwaarden

University of Colorado at Boulder

In partial fulfillment of the requirements for EDUC 8720

Dr. Derek Briggs

May 3, In Press

Abstract

This simulation study examines Rasch model mean-square fit statistics under several small-sample (e.g. pilot test) conditions and highlights considerations that should be taken. Samples smaller than 100 are largely ignored in the extant literature. As a result, it is unclear whether the same guidelines for their use apply or whether there are unique limitations of these residuals-based fit statistics in these situations. In order to evaluate these issues, dichotomous item response data are generated and calibrated from samples of 30, 50 and 70 respondents with both homogenous and heterogeneous ability levels. Test construction is designed to expect varying degrees of misfit in order to determine Type I error rates and statistical power.

Introduction

It is often noted that modeling item response data with the one-parameter logistic (1PL or "Rasch") model can potentially provide the necessary properties for objective measurement. Many of these properties, such as specific objectivity (or the ability to evaluate two test takers independently of the items used to compare them) or conjoint additivity, are necessary for the fundamental measurement of a latent trait and the establishment of interval scales (e.g. Bond & Fox, 2007; Rasch, 1977; Wright, 199X). An essential corollary to this claim is that it must be shown that the model fits the data (or vice versa) in order for these properties to hold: they are not obtained simply through the application of the model to the data on hand (Borsboom, 2004; Wu & Adams, In Press). Important decisions must be made when data generated from items does not conform to the model, such as whether item response data should be excluded from an analysis, an item should be removed from a test, or an alternative model that fits the data better (but does not offer the same statistical benefits as the 1PL model) should be used.

These critical decisions suggest the need for good indicators of whether the data are adequately consistent with the model. Indeed, several statistical tests of model fit are available for this general purpose, and they address varying aspects of model fit (Fischer & Molenaar, 1995, Wu & Adams, In Press). For example, some tests evaluate particular item level assumptions such as equal item discrimination, unidimensionality, and local independence. Other statistics evaluate the overall model fit at either the test or item level. One major distinction between fit statistics is between those that are based on ability group-level differences between the expected and observed number responses and those that use standardized residuals based on individuals' observed responses. This study is focused on the later type, which follow from the residual based statistics that Wright and Panchapakesan (1969) first proposed for the

analysis of Rasch model fit. This particular type of fit statistic is widely recommended in the literature (Bond & Fox, 2007; Smith, 1990; Wilson, 2XXX; Wu & Adams, In Press) despite others' arguments against their use (Divgi, 1986; George, 1979; Karabatsos, 2000).

Much of the debate about whether or not the use of residual-based fit statistics is appropriate returns us to the concept that we need good indicators of whether the data are adequately consistent with the model. This need requires a definition of a "good indicator" and what signifies "adequately consistent". From a statistical standpoint, any good test requires that the statistic's distributional properties are well understood. Knowledge about the expected values and variation of a particular statistic's distribution (specifically the "null" distribution of the statistic when the data is known to fit the model) allow us to make judgments about whether observed data is likely to have been generated through a process that the model adequately describes. That is, it is possible to establish critical values that correspond with the level of expected Type I errors (i.e. rejecting the null hypothesis that the data fit the model when the null is actually true). A good indicator should also provide adequate power to detect misfit when it does exist (i.e. to reject the null when it is in fact false, avoiding type II errors).

Many authors have addressed the issue of whether residual fit statistics are "good indicators" that are capable of providing trustworthy information about model fit through simulation studies. Paradoxically, some have concluded that their use is justified (e.g. Smith, 1990; Wright & Masters, 1982; Wu & Adams, In Press), while others have found them to be inadequate or inappropriate (e.g. Divgi, 1986; George, 1979; Karabatsos, 2000). Regardless of their ultimate conclusions, a common thread throughout is that these studies are generally focused on the use of these statistics in relatively large samples (usually greater than 100 respondents). Only Smith (1991) includes a small sample (n = 30) in his analysis, but this

particular condition is not the focus of his analysis. Subsequently, the assertions made about the distributional properties of residual fit statistics have only been shown to hold asymptotically (i.e. they are known to hold only as the number of respondents increase, approaching infinite size). Researchers and practitioners applying the Rasch model to small-sample data are subsequently left to speculate about (or unwittingly ignore) the realities about the distributions of these fit statistics and the appropriate use of them in determining whether the model is consistent with their data. [WU AND ADAMS + WC = USE T STATS]

This is problematic because decisions about whether to include an item in a test or survey instrument are often times made based on data from a small sample (e.g. a pilot test situation). The potential to make misinformed decisions in conditions often encountered in these situations (such as having a sample with ability levels that are unrepresentative of the population's, or lacking full overlap of ability/item difficulty) is substantial given the present lack of guidance or caution offered for the use and interpretation of fit statistics. The aim of this study is to address this gap in the literature by assessing Type I error and statistical power of residual-based fit statistics (including their transformations to unit normal statistics) in samples with fewer than 100 respondents and under various, likely unfavorable, conditions.

In addressing this topic, I begin by first presenting the formulae used to derive the standardized residuals and their associated fit statistics while also describing their origins, criticisms that have been raised and approaches taken to address those issues. I also detail the function of the fit statistics, describing the purposes for which they can be used and the factors that influence their distribution. Subsequently, I detail the methods I use to evaluate the small-sample properties of the fit statistics, including the data simulation, analysis and transformation processes. Finally, I present the results of the study, lessons learned and its limitations.

Background

The Rasch model makes plausible assumptions about what occurs when a person responds to an item. The purpose of item fit statistics is to evaluate these assumptions and detect items that produce response data that are inconsistent with the applied model (Wright & Masters, 1982; Wright & Stone, 1979). It is important that fit statistics provide accurate, reliable information in order to correctly identify items that should be considered for removal from a particular test instrument (or flagged for revision) or whether data generated from a problematic item should be excluded from an analysis. This study evaluates the family of χ^2 residual-based item fit statistics that Wright and Panchapakesan (1969) proposed, as well as transformations of them to standard-normal deviates (Wright & Masters, 1982; Wright & Stone, 1979; Wu & Adams, In Press) and corrections to those transformations (Smith, 1991; Wang & Chen, 2005). Because there are several versions of these statistics, their transformations and corrections, it is helpful to retrace their history and highlight the reasons provided for them.

Although the current manifestation of standardized residual fit statistics are based on individual respondents, distinct from other fit statistics that group respondents by ability level, the original formula to calculate a "residual" categorizes respondents by raw score and correct item response groups (Smith, 1991). The original "standard deviate" residual, Z_{ij}, Wright and Panchapakesan (1969) proposed is $Z_{ij} = \frac{a_{ij} - r_j p_{ij}}{[r_j p_{ij}(1 - p_{ij})]^{1/2}}$, where a_{ij} represents the observed number of correct responses for persons with raw score j, r_j the number of persons with raw score j, and p_{ij} is the probability of a correct response on that item for score group j. This probability is assumed to be a function of the interaction between a respondent's ability, b_j, and the item difficulty, d_i, such that $p_{ij} = \exp(b_j - d_i)/(1 + \exp(b_j - d_i))$. It is important to note that the values of b_j and d_i are estimated from the data. Consequently, this statistic is only an

approximate chi-square (Wright & Stone, 1979; Smith, 1991). The corresponding squared standardized residuals (for the entire data matrix) are $\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{m} Z_{ij}^2$ with (L-1)(m-1) degrees of freedom, where L is the number of items and m represents the number of groups with a unique raw score, and $r_j \neq 0$. Wright and Panchapakesan (1969) also proposed a fit statistic for the individual items: $\chi_i^2 = \sum_{j=1}^{m} Z_{ij}^2$ with m-1 degrees of freedom.

Subsequent versions of the standardized residuals move away from the ability group residual, using an item-person residual instead (Wright, 1977; Wright & Stone, 1979). Here, $z_{ni} = \frac{x_{ni} - p_{ni}}{(w_{ni})^{1/2}}$ where $x_{ni} - p_{ni}$ is the simple residual (the probability of a correct response for person i subtracted from their observed score on item n) and $w_{ni} = p_{ni}(1 - p_{ni})$ is the variance of x_{ni}, which is the variance of the well known Bernoulli distribution. These residuals (both group and individual based) are distributed approximately chi-square with one degree of freedom (Smith, 1988; Wright & Panchapakesan, 1969; Wright & Stone, 1979; Wu & Adams, In Press).

This standardized residual can be used to form a mean square fit statistic by summing over all respondents and dividing by the sample size: $u_i = \sum_{n=1}^{N}(z_{ni})^2 / \quad N$

In order to form critical values for expected Type I errors, it is necessary to establish the expected variability in the statistic's distribution. Smith (1991) provides the standard deviation $SD_{u_i} = [\sum_{n=1}^{N}(1/w_{ni}) - 4N]^{1/2} / \quad N$. Other authors suggest approximations of the mean square variance. Examples are $VAR_{u_i} = \sum_{n=1}^{N}(c_{ni}/w_{ni}^2)/N^2 - \frac{1}{N}$ where c_{ni} is the kurtosis of x_{ni} (Wang & Chen, 2005; Wright & Masters, 1982), and the use of 2/N is suggested as an asymptotic approximation that is estimated from the sample size and not the mean-square statistic itself (Wu

& Adams, In Press). Smith (1991) does not provide a derivation for his suggested standard deviation, and the derivation of the variance, s^2, has only recently been documented in the literature (Wu & Adams, In Press). However, as I describe in the results of this study below, the choice of formula can produce very different estimates of the mean-square variance in small samples. This translates to differences in the unit transformations as well.

Regardless of the formula(s) used, it is clear that the value of the mean square variance is dependent upon the sample size, which makes the establishment of set critical values impossible (Smith, 1991; Wang & Chen, 2005; Wu & Adams, In Press). Despite this fact, many other authors still suggest the use of such critical values (Bond & Fox, 2007; Wilson, 2XXX; Wright, Linacre, Gustofson, & Martin-Lof, 1994). Specifically, they suggest the range of .77 to 1.3, which was originally suggested in a software manual (Adams & Khoo, 1996).

Wright and Stone (1979) avoid the issues of establishing the variance of the mean square statistic and setting specific critical values, and instead suggest that it is F distributed (when the summed standardized residuals are divided by N-1), which "can conveniently be evaluated as the t-statistic" (Wright & Stone, 1979, p. 77) with an approximate unit normal distribution using the transformation: $u_i^* = \sum_{n=1}^{N} z_{ni}^2 / f \quad and \quad f = N - 1,$, such that $t_{u_i} = [\ln(u_{ni}^*) + u_{ni}^* - 1][f/8]^{1/2}$. This particular *t*-transformation, however, is only one of the suggested transformations. Several other authors present the cube root transformation when an estimate of the mean square variance is available: $t_{u_i} = (u_i^{1/3} - 1)(3/SD_{u_i}) + (SD_{u_i}/3)$, (Wright & Masters, 1982; Wang & Chen, 2005), or a Wilson-Hilferty cube root transformation, $t_{u_i} = (u_i^{1/3} - 1 + 2/(9N)) + (2/(9N))^{1/2}$, as an asymptotic approximation when the variance is unavailable (Wu & Adams, In Press).

In addition to the discrepancies in the suggestions of how to calculate the variance of the mean square statistic and transform them into a unit normal distribution, there is also

disagreement over whether the standardized residuals can be treated as a normal variate. This has subsequent implications for how the distributions of the mean square statistics and their derivatives are treated. George (1979) and Divgi (1986) both object to their use on the grounds that the residuals come from a discrete variable and are therefore discrete themselves, meaning that they cannot be truly normally distributed and their squares cannot be chi-square distributed. In short, the assumption invokes the normal approximation of binomial variables, which is admissible in large samples. However, in this case the sample size is one (for each person and item interaction) and consequently error is introduced when the residuals are treated as though it were sampled from a normal distribution (George, 1979) and tests based on them can produce misleading results (Divgi, 1986).

There still does not seem to be any general agreement about the distributional form of either the residuals or the fit statistics or what might be a reasonable approximation. Many authors continue to accept the normality assumption despite its flaws, arguing that they provide useful results when samples are large enough and items are plentiful (Smith, 1991; Wang & Chen, 2005; Wu & Adams, In Press). What is now more universally agreed upon is the fact that sample size must be taken into account when critical values of the untransformed mean square statistics are established, or a standardized transformation can be used for any sample size (Wang & Chen, 2005; Wu & Adams, In Press). However, the transformed values that IRT software provide (or the researcher constructs) will depend upon the particular formula used for the variance and the transformation itself, as described above. Although these values likely converge to similar conclusions as sample and item numbers increase (Wu & Adams, In Press), this is potentially more problematic for small samples.

Furthermore, transformed values of the residual mean square fit statistics have repeatedly shown mean values significantly less than zero and variances less than one in simulation studies, which requires a correction to be applied (Smith, 1991; Wang & Chen, 2005). Wright and Masters (1982) addressed this early on saying, "We have not mastered the statistical details of [the departures from expectations and variances] well enough to provide useful corrections for reducing them. But we have not found this an impediment to practice" (p. 101). However, it is not clear whether this subjective analysis is true in all situations, and many researchers seem to be unaware of this situation or ignore it (Wilson (2XXX) and Wu and Adams (In Press), for example) and software documentation rarely specifies whether or not any corrections are made to the transformations (if they are provided at all).

In their discussions of possible transformation correction measures, both Smith (1991) and Wang and Chen (2005) note that these problems occur to a greater extent in items that have more extreme difficulty values (i.e. they are located further away from the mean sample ability). Smith (1991) suggests that this occurs because item difficulty and person ability are estimated together, and those estimates are subsequently used in the calculation of the residuals. This implies a lack of independence between the residuals, which is another criticism of the assumption that the residuals are distributed chi-square (independence of the squared variables is required). Accordingly, corrections account for both person ability and item difficulty. Smith (1991) also suggests that test length is an influential factor. The corrections to the means and standard deviations that he suggests are $C_{Mean(u_i)} = \frac{L^2 + ID^2}{L^2 - 5}$, where L is the number of items, N is the sample size and $ID = ((\Sigma|b_n - d|)/N)^2$ with b_n representing the ability of person n and d is the difficulty of the item. The correction for the standard deviation is

$C_{SD(u_i)} = \left|\frac{NL}{(N-1)(L-2)}\right| \left|1+\frac{ID^2}{L^2}\right|$. Notably, this correction is applied to the calculated mean-square statistic and standard deviation and subsequently used in a transformation formula. Also, the later corrections are not applied directly to the observed mean square, but rather its inverse is multiplied by the estimated standard deviation (Smith, 1991). This is in contrast to the correction Wang and Chen (2005) provide, which is applied directly to the cube root transformed statistic or as a modification to the acceptable critical range of values. Here $C(t_i.CR) = \frac{t_i.CR}{1-\left|\bar{b}-d\right|/8}$ & $C(CritRange_{t_i.CR}) = CritRange_{t_i.CR} \times (1-\left|\bar{b}-d\right|/8)$, where $\bar{b}$ is the mean person ability and *d* is the estimated item difficulty.

The weighted mean-square statistic, or INFIT , is another attempted "correction" to the distribution of the unweighted mean-square statistic (OUTFIT) that takes the relationship between estimated respondent ability and item difficulty into account. This statistic is more documented, and controversial, than the corrections just described. The names given to these statistics are based on the item and person types that most influence them. For the unweighted OUTFIT statistic, the "out" comes from it being "outlier" sensitive. That is, it is heavily influenced by unexpected observations at the tails of the distributions (i.e. correct responses to difficult items from respondents of low abilities and high ability respondents' incorrect responses to easy items). Wright and Masters (1982) suggested that one might mistakenly "reject an item as misfitting because of just two or three surprising responses made by persons for whom the item was quite inappropriate" (p. 99). This remark suggests problems with the OUTFIT statistic in small sample (e.g. pilot test) situations. In large samples, it would be expected that there would be relatively few of these unexpected responses. If many are observed, then there is reason to suspect an issue with the item and the OUTFIT statistic is performing as expected.

However, continuing with this implicit small sample context, Wright and Masters (1982) suggest that "[a]n alternative is to weigh the squared residuals so that responses made by persons for whom the item is remote have less influence on the magnitude of the item fit statistic" (p. 99). This is supposed to more heavily weight the items that provide the most information (and thus the "IN" in INFIT) about a respondent. Although this may make sense in a small sample context, where a handful of unexpected responses can potentially have a sizable impact, the use of the INFIT statistic has also been criticized because unexpected responses may be the ones we are most interested in, and weighting them out of the equation (in large samples) is effectively "cooking the books" in favor of the items, when if fact the items may perform poorly (CITE). Regardless of this viewpoint, they are of interest here given their potential utility in small sample situations.

The weighted mean square is calculated as $v_i = \sum_{n=1}^{N}(x_{ni} - p_{ni})^2 / \sum_{n=1}^{N} w_{ni}$ and the variance is given as $VAR_{v_i} \sum_{n=1}^{N}(c_{ni} - w_{ni}^2)/(\sum_{n=1}^{N} w_{ni})^2$ in Wright and Masters (1982) or the standard deviation $SD_{v_i} = [\sum_{n=1}^{N} w_{ni} - 4\sum_{n=1}^{N} w_{ni}{}^2]^{1/2} / \sum_{n=1}^{N} w_{ni}$ can be used (Smith, 1991). The same formulae for the cube root transformation to a standard-normal deviate can be used, substituting either the approximated INFIT standard deviation or the square root of its variance. Smith (1991) and Wang and Chen (2005) also provide corrections specific to INFIT . Respectively, they are $C_{Mean(v_i)} = \frac{L^2 + ID^2}{L^2 - 1}$ and $C_{SD(v_i)} = \left|\frac{NL}{(N-1)(L-1)}\right| \left|1 + \frac{ID^2}{L^2}\right|$, and

$$C(t_i.CR.IN) = \frac{t_i.CR.IN}{1 - |\bar{b} - d|/4} \quad \& \quad C(CritRange_{t_i.CR.IN}) = CritRange_{t_i.CR.IN} \times (1 - |\bar{b} - d|/4).$$

The statistical framework supporting mean square fit statistics seems shaky at best, often propped up by approximations to normality and the unsatisfactory (at least in the small-sample

context of this study) buttress of asymptotic properties. There is a lack of consistency in the field as far as how to treat and derive these statistics, what their properties are and the extent to which they are ultimately useful and dependable. At the end of the day, it is suspicious that these statistics require such extensive transformation and correction in order to be well behaved: their use in any context, but particularly in small samples, seem quite perilous indeed.

Methods

In this study I use data simulation to shed light on the null distribution of mean-squared residual fit statistics and its derivatives. Although the analysis of observed data is the ultimate purpose of any fit statistic, real-world data are unhelpful in the evaluation of a test statistic's performance because of the uncertainty inherent in the data: the items and instruments are unproven and so the characteristics of the data they generate are unknown, which is why we need fit statistics in the first place. Conversely, simulated data provide a verifiable proving ground in which to engage the statistics under controlled conditions. When the performance of fit statistics is well understood, they effectively "provide a frame of reference against which the performance of a given item or person can be judged" (Smith, 1991, p. 547). Simulation studies are an excellent way of establishing this frame, which include the expected values, variances, and Type I error rates when the data is simulated to fit the model. The statistical power to detect abnormalities can be evaluated when the data is simulated to misfit the model in particular ways. The researcher is only limited by the choices made about which model features to examine and which data deviations to focus on. In the following I describe the choices I made for this study.

Simulation. Item response data are generated from 20-item tests given to 1000 samples of respondents under each of 6 conditions. The first set of conditions is test type: half of the tests contain items that conform to the one-parameter logistic (1PL) model while the other does not,

specifically due to varying item discriminations. The former test type produces data that will fit the model, which provides information about the expected values and variances, as well as Type I error rates (false positive indications of misfit). The latter half will produce misfit data, which provides information about the Type II error rates (false negatives, or failure to indicate misfit when present), and are specifically designed to evaluate deviations in discrimination because residual fit statistics evaluate the consistency of the slope parameter across items (Wu & Adams, In Press). Here I use three different slope parameters (-.5, 1, 1.5)[1] to evaluate whether the test statistics correctly identify items that are more or less discriminating than the average item, which is specified to equal 1. This test type can be seen as equivalent to a two-parameter logistic (2PL) model or as three 1PL model sub-tests nested within a large one. The 20 item difficulty parameters are specifically set to range from -1.8 to +2 logits in increments of .2 logits (rather than randomly selecting them from a probability distribution as is commonly done in some simulation studies). Specifically setting the item parameters and simulating the test 1000 times is equivalent to a thought experiment in which a test of known properties is repeatedly administering to many different samples of similar subjects. The randomness in the experiment is thus constrained to come only from the randomly generated ability distribution of respondents and the interaction between them and the set items.

The second simulation condition is the specification of heterogeneous and homogenous respondent ability distributions. The former group is selected from a unit normal distribution (mean = 0, standard deviation = 1.0) and the latter group is selected from a non-central normal with a restricted variability (mean = 0.75, standard deviation = 0.5). A heterogeneous ability distribution satisfies the condition that all items have an adequate number of respondents across the range of item difficulties to accurately evaluate the item. Conversely, item responses from a

[1] More extreme deviations were also explored. The results were very similar and not presented here.

homogenous ability distribution (disproportionately above average in this case) could have a higher Type I error rate for the items with poor coverage (the easy items).

The final condition variable is sample size. Data is generated for both test types and ability distribution combinations for samples of 30, 50 and 70 respondents. This shows the effect that relatively small sample sizes (typical of what one might find in pilot test situations) may have on the fit statics' performance and their distributions. Importantly, the null distribution of these fit statistics will necessarily vary for any combination of item number, sample size and ability distribution in relation to item difficulty (Smith, 1988; Smith, 1991). Thus the results from this study can only be generalized to the six conditions similar to those simulated here.

Calibration. The data generation processes is carried out in `R`, a language and environment for statistical computing (R Development Core Team, 2009). The item calibration process is performed in tandem with each sample generation using the `eRm` () package. Although the combination of these two processes greatly increase the speed and efficiency of the simulation process, allowing me to specify a greater number of replications and conditions, it also presents new issues that must be addressed. For example, the `eRm` program estimates person and item parameters via conditional maximum likelihood (CML), which allows respondent ability and item characteristics to be estimated independently of each other, and does not make any a priori assumptions about the distribution of student abilities, as is done in the more commonly used marginal maximum likelihood (MML) estimation approach. This is potentially very beneficial as it has been noted that some of the issues with residual based fit statistics, such as the restrictions in the means and variances, may in part be due to the dependence between items and persons that is present when they are estimated together (Smith, 1991). The assumptions about sample ability distribution made under MML may also bias these

statistics' estimates, particularly when sample and item numbers are small, giving the prior distribution a greater amount of influence on the estimates.

The use of CML estimation in this study potentially restricts the findings because most IRT software packages use MML estimation. The results of this study may be misleading (or at least not universally relevant) if substantial differences do indeed exist between the fit statistics produced under the two estimation approaches. I attempted to address this issue through the with an analysis of the first 50 data sets generated under each condition using the commercial IRT software program Conquest. However, the program provided aberrant results for the fit statistics produced during repeated estimation runs on the same data set. Therefore any comparisons of MML and CML estimation approaches are not presented here, but this issue remains an active area of my current research.

Another issue presented from the use of the `eRm` package is that the untransformed mean-square INFIT and OUTFIT are the only fit statistics the package provides. However, the matrix of expected values (i.e. probabilities of a correct response) is made available. I use this matrix to back out the standard deviation and kurtosis of the residuals using the various formulas presented above, and then compute necessary variances and standard deviations in order to construct the *t*-transformations and corrections. The performances of the fit statistics variants are then compared with one another in the various small sample conditions.

Results

Perhaps one of the most intriguing results from this study is became immediately apparent in the analysis of Type I error under the null distribution conditions: because the distribution of the fit statistics is dependent upon the distribution of respondent ability and item difficulty, the Type I error rates under the null distribution are also dependent upon these factors.

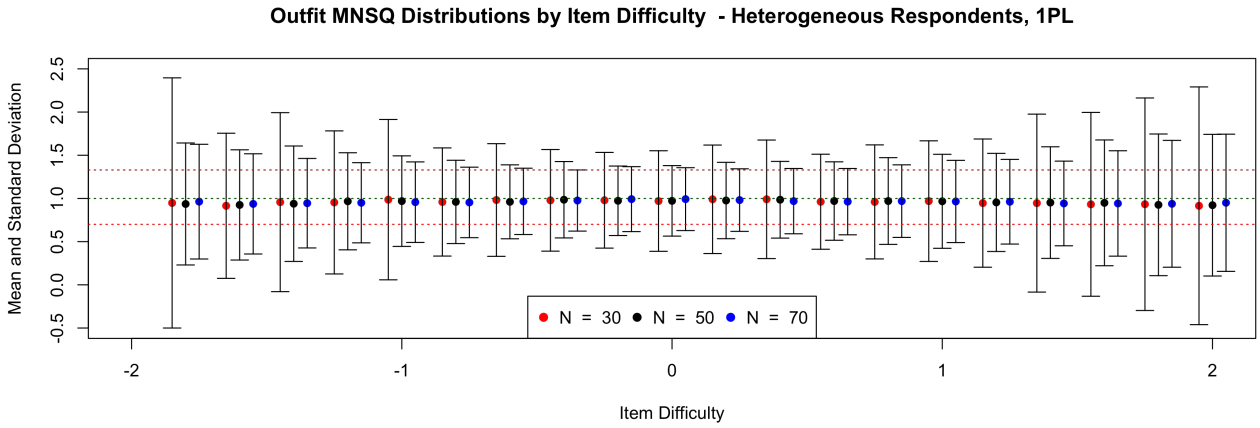
In past simulation studies, summaries of the null distribution, including the distribution moments and the rates of false-positive indications of misfit, are exclusively reported in aggregate over all items in a test (c.f., Smith, 1988; Smith, 1991; Wang & Chen, 2005; Karabatsos???). These presentations are misleading in that one can believe that there is only one Type I error for a particular combination of items and respondents, when in fact these rates are item specific. The extent to which this holds for large-samples and varying test sizes (or the use of MML estimation) is not evaluated here. These factors could dramatically reduce this effect, justifying the item aggregation to the test level. However, it is noteworthy in the small-sample conditions examined here, and therefore all results are presented at the item level.

Null Distribution and Type I error rates. Table 1 summarizes the OUTFIT mean-square null distribution over the three sample sizes. The mean value departs from the expected value of 1 to a greater extent as you move towards the item difficulty extremes, and the variability increases as well. Figure 1 provides this same information graphically, with error bars extending to 2 standard deviations from the mean, and the traditional fixed critical values (1.3 and 0.77) are displayed as red dotted lines.

Table 1. Means and Standard Deviations of Mean Square OUTFIT Statistics

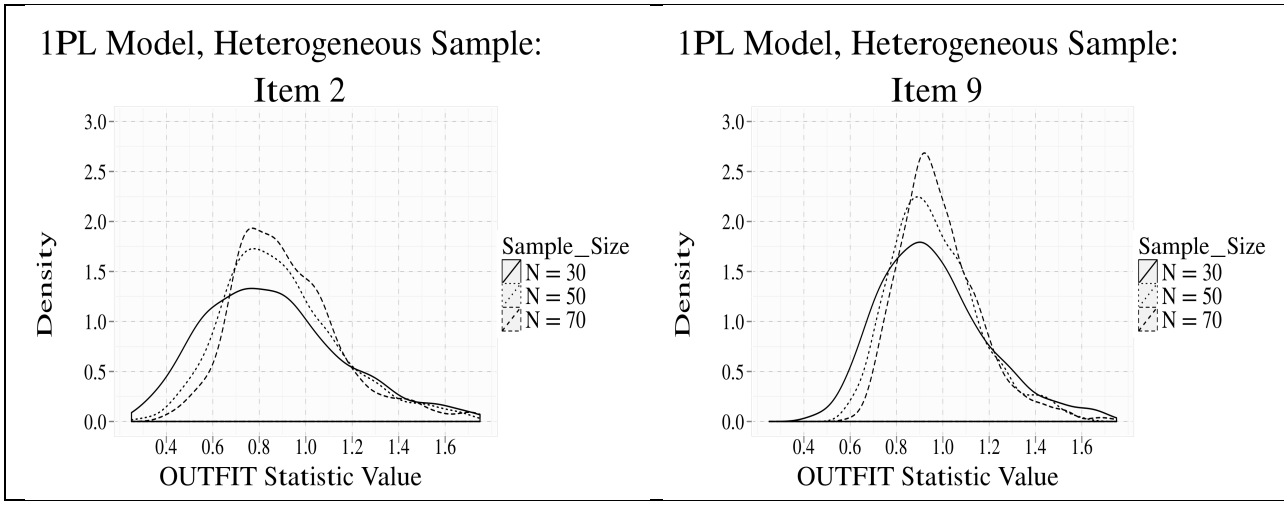
		N = 30		***N = 50***		***N = 70***	
	Delta	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Item 1	**-1.8**	0.95	*0.72*	0.94	*0.35*	0.96	*0.33*
Item 2	**-1.6**	0.92	*0.42*	0.93	*0.32*	0.94	*0.29*
Item 3	**-1.4**	0.96	*0.52*	0.94	*0.33*	0.95	*0.26*
Item 4	**-1.2**	0.95	*0.41*	0.97	*0.28*	0.95	*0.23*
Item 5	**-1.0**	0.99	*0.46*	0.97	*0.26*	0.96	*0.23*
Item 6	**-0.8**	0.96	*0.31*	0.96	*0.24*	0.95	*0.20*
Item 7	**-0.6**	0.98	*0.33*	0.96	*0.21*	0.97	*0.19*
Item 8	**-0.4**	0.98	*0.29*	0.99	*0.22*	0.98	*0.18*
Item 9	**-0.2**	0.98	*0.28*	0.97	*0.20*	0.99	*0.19*
Item 10	**0.0**	0.97	*0.29*	0.97	*0.20*	0.99	*0.18*
Item 11	**0.2**	0.99	*0.31*	0.98	*0.22*	0.98	*0.18*
Item 12	**0.4**	0.99	*0.34*	0.99	*0.22*	0.97	*0.19*
Item 13	**0.6**	0.96	*0.28*	0.97	*0.23*	0.96	*0.19*

Item 14	**0.8**	0.96	*0.33*	0.97	*0.25*	0.97	*0.21*
Item 15	**1.0**	0.97	*0.35*	0.97	*0.27*	0.97	*0.24*
Item 16	**1.2**	0.95	*0.37*	0.95	*0.28*	0.96	*0.25*
Item 17	**1.4**	0.95	*0.52*	0.95	*0.32*	0.94	*0.25*
Item 18	**1.6**	0.93	*0.53*	0.95	*0.36*	0.94	*0.31*
Item 19	**1.8**	0.93	*0.62*	0.93	*0.41*	0.94	*0.37*
Item 20	**2.0**	0.92	*0.69*	0.92	*0.41*	0.95	*0.40*



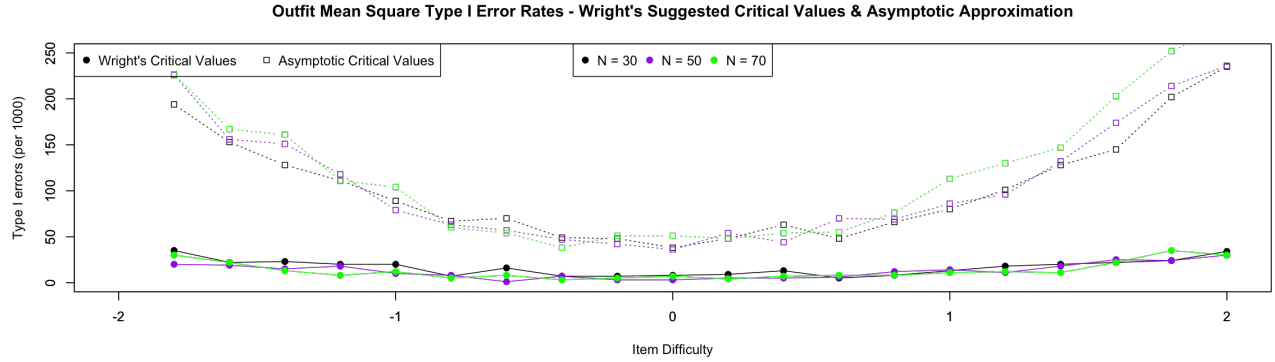
The picture provided in Figure 1 can be misleading. It could be interpreted that the distributions of these statistics are symmetrical about the mean, which is not the case. Figure 2 below shows the distribution of the OUTFIT statistic for two items and all sample sizes. These plots show that the skewness and kurtosis of the distributions is also affected by the location of the item. Item 18, whose generating difficulty parameter, δ_i, is set to +1.6 is a near copy of item 2 with a difficulty of -1.6. It can be seen from these examples that although the average fit statistic may be near the expected value of 1 the bulk of them will be below that, calling into question the use of critical values that are equidistant from 1.

Figure 2.



Returning focus to Figure 1, one notices that the variance of the distributions decrease quickly as the sample size increases from 30 to 70, but that the traditional set critical values remain within 2 SDs of the majority of item distributions. Thus in small sample situations, one would have an unacceptably high level of Type I error using this criteria. However, two other criteria that take sample size into account have been discussed in the literature and can be used as alternatives. The first is the use of $1\pm6/N^{1/2}$, which Ben Wright suggested to Smith and his colleagues (1998) in a personal communication. The second uses the asymptotic approximation of the variance: $1\pm(2/N)^{1/2}$ (Wu & Adams, In Press). Figure 2 shows the Type I error rates across items using these two criteria. The use of asymptotic critical values is obviously inappropriate here (as would be expected in small samples), particularly as the difference between the item difficulty and the average ability level, 0 logits, increases. Curiously, the overall Type I error rate increases from 10% to 12% as sample size increases from 30 to 70 (this may be due to sampling variation and not a sign of increasing Type I error as sample size increases). Although it is less noticeable in comparison to the asymptotic critical values, the Type I error rates using Wright's suggested values displays a similar convex shape, as depicted in the second panel.

Figure 3.



Importantly, these indications of misfit do not occur at the high and low ends of the critical values (i.e. positive and negative OUTFIT) in equal proportions. In fact, it is impossible to have a statistic less than the lower critical value using Wright's suggested formula at a sample size of 30 (it is a negative number). The Type I error rates (which average between 1% and 2% across all items) are thus restricted to indications of positive OUTFIT: the tests of misfit are reduced to a one-tailed test. Although some authors have suggested that items with low fit statistic values indicate more discriminating (and therefore "better") items, which can be ignored, these items would still indicate a failure of the Rasch model to fit the data. If it is impossible to indentify this type of item with this statistic and critical values, then it is ultimately inadvisable to use them in small samples.

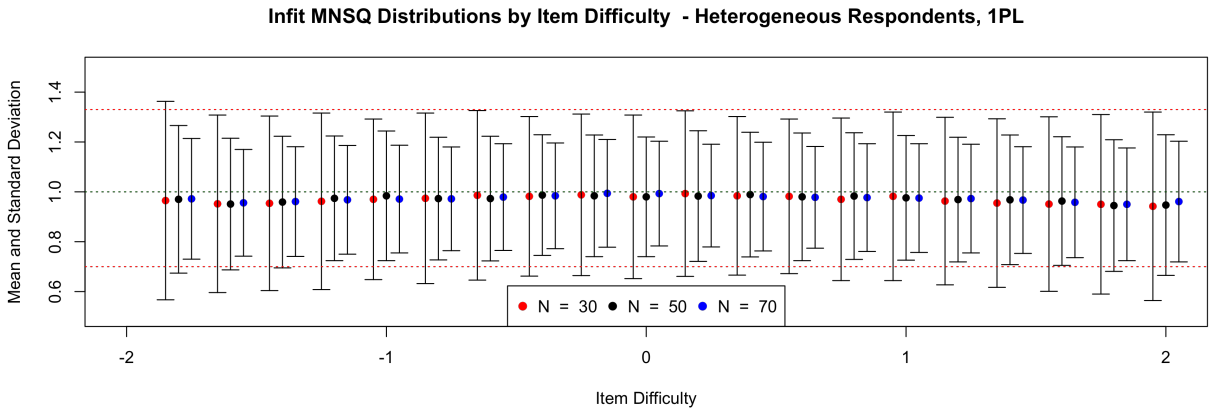
It is then necessary to evaluate the INFIT statistic null distribution and Type I error rates to see whether these problems exist here as well. Table 2 shows that a similar patter in the

means and standard deviations is present, but to a lesser extent. Figure 4 presents the same information graphically. However, here we see that the two standard deviation error bars are well within conventional critical values once the sample size increases beyond 30, suggesting that very few items would be rejected using these criteria.

Table 2. Means and Standard Deviations of Mean Square OUTFIT Statistics

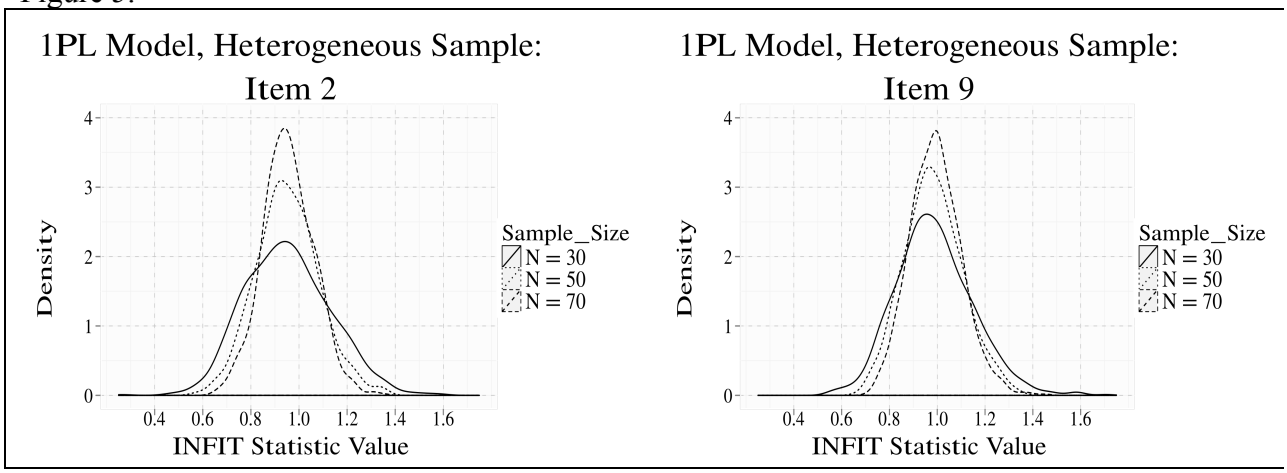
		N = 30		***N = 50***		***N = 70***	
	Delta	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Item 1	**-1.8**	0.97	*0.20*	0.97	*0.15*	0.97	*0.12*
Item 2	**-1.6**	0.95	*0.18*	0.95	*0.13*	0.96	*0.11*
Item 3	**-1.4**	0.95	*0.18*	0.96	*0.13*	0.96	*0.11*
Item 4	**-1.2**	0.96	*0.18*	0.97	*0.13*	0.97	*0.11*
Item 5	**-1.0**	0.97	*0.16*	0.98	*0.13*	0.97	*0.11*
Item 6	**-0.8**	0.97	*0.17*	0.97	*0.12*	0.97	*0.10*
Item 7	**-0.6**	0.99	*0.17*	0.97	*0.13*	0.98	*0.11*
Item 8	**-0.4**	0.98	*0.16*	0.99	*0.12*	0.98	*0.11*
Item 9	**-0.2**	0.99	*0.16*	0.98	*0.12*	0.99	*0.11*
Item 10	**0.0**	0.98	*0.16*	0.98	*0.12*	0.99	*0.11*
Item 11	**0.2**	0.99	*0.17*	0.98	*0.13*	0.99	*0.10*
Item 12	**0.4**	0.98	*0.16*	0.99	*0.13*	0.98	*0.11*
Item 13	**0.6**	0.98	*0.16*	0.98	*0.13*	0.98	*0.10*
Item 14	**0.8**	0.97	*0.16*	0.98	*0.13*	0.98	*0.11*
Item 15	**1.0**	0.98	*0.17*	0.98	*0.13*	0.98	*0.11*
Item 16	**1.2**	0.96	*0.17*	0.97	*0.13*	0.97	*0.11*
Item 17	**1.4**	0.96	*0.17*	0.97	*0.13*	0.97	*0.11*
Item 18	**1.6**	0.95	*0.18*	0.96	*0.13*	0.96	*0.11*
Item 19	**1.8**	0.95	*0.18*	0.95	*0.13*	0.95	*0.11*
Item 20	**2.0**	0.94	*0.19*	0.95	*0.14*	0.96	*0.12*

Figure 4.



The shape of the INFIT distribution is also far more symmetric than the OUTFIT statistic, although it is still clearly skewed, favoring values lower than expected as shown in Figure 5.

Figure 5.

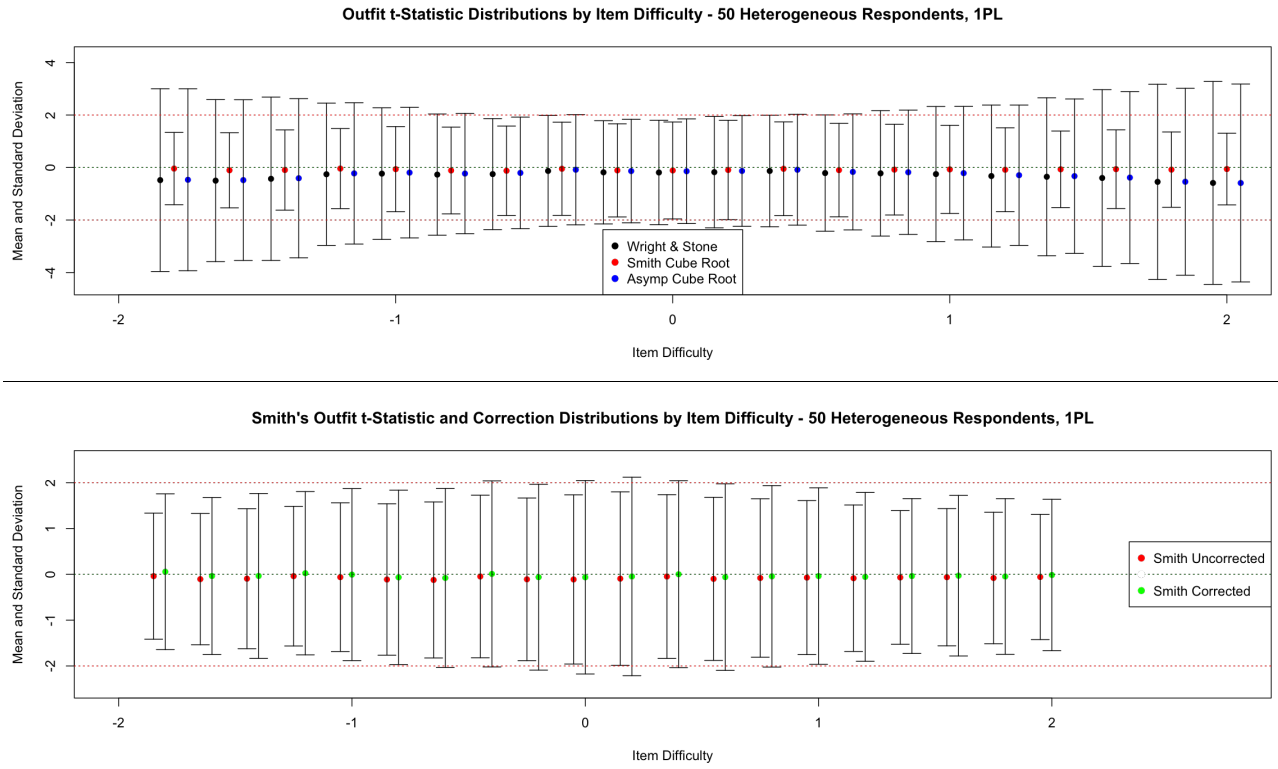


I use Ben Wright's suggested critical values for INFIT, $1\pm2/N^{1/2}$, (Smith, et al., 1998) to evaluate the INFIT Type I errors (Wu and Adams (2010) do not provide an asymptotic approximation). The results are presented graphically in Figure 6 below. The overall Type I error rate for the INFIT statistic is 3% at all three sample sizes. Positive and negative misfit indications are evenly distributed for the sample size of 30, but disproportionately more negative misfit is indicated as *N* increases. Overall the INFIT statistic appears to have a more consistent and even distribution than OUTFIT in small sample sizes.

Figure 6.

Evaluation of the distributions and Type I errors for the unit-normal transformation of OUTFIT and INFIT statistics is convoluted given the various transformations and corrections discussed above. Graphical representations provide a basic picture of this and are presented below. Tables of the values used to produce these plots are provided in the Appendix. The top panel of Figure 7 shows the distribution of three transformations, including the log transformation proposed by Wright and Stone (1979), the cube root transformation using the calculation of the OUTFIT standard deviation proposed by Smith (c.f., 1991), and the Wilson-Hilferty cube root asymptotic approximation (Wu & Adams, In Press). Missing from this picture is the cube root transformation using the calculation of the OUTFIT variance Wright and Masters (1982) propose. The calculated variances were quite large and the resulting error bands were quite large, rendering them un-useful. This figure is also restricted to a sample size of 50, although the picture is very much the same for the two other samples. The hour-glass figure becomes slightly less exaggerated as sample size increases, but not dramatically. Interestingly, the logarithmic and asymptotic transformations are nearly identical (correlated at about .98 across all sample sizes).

Figure 7.



The bottom panel of Figure 7 displays Smith's transformation at a sample size of 50 again along with the correction. Here we see that Smith's transformation does not follow the typical hour-glass shape, and is actually restricted at the tails of the item difficulty extremes. The correction helps improve this overall, although not entirely.

Figure 8 shows the Type I error rates associated with the transformations. Each follows the distribution discussed above, as expected. The Type I errors are disproportionately dispersed between high and low values as well. The Wright and Stone and asymptotic transformations have a disproportionate number of Type I errors below 2 standard deviations, but a fairly consistent indication of high values (averaging 5% across items). Smith's transformation has more false positive indications in the middle of the item difficulty spectrum, but never surpasses 5% for any one item. However, nearly all of these indications of misfit are for values above the 2 SD critical value (i.e. positive misfit).

Figure 8.

Far fewer transformations have been offered for the INFIT statistic than for OUTFIT, and the variance calculation for the Wright and Masters (1982) transformation is also too large to be useful here. As a result, only the distribution of Smith's transformation and correction are presented in Figure 9. The restriction of the variance at the ends of the item difficulty distribution is even more noticeable here. As a result, the Type I error for this correction is disproportionately centered on the items near the average of the distribution. However, the false positive indication for any one item never surpasses 5%. Figure 10 displays this as well as the error rate for the Wright and Masters transformation

Figure 9.

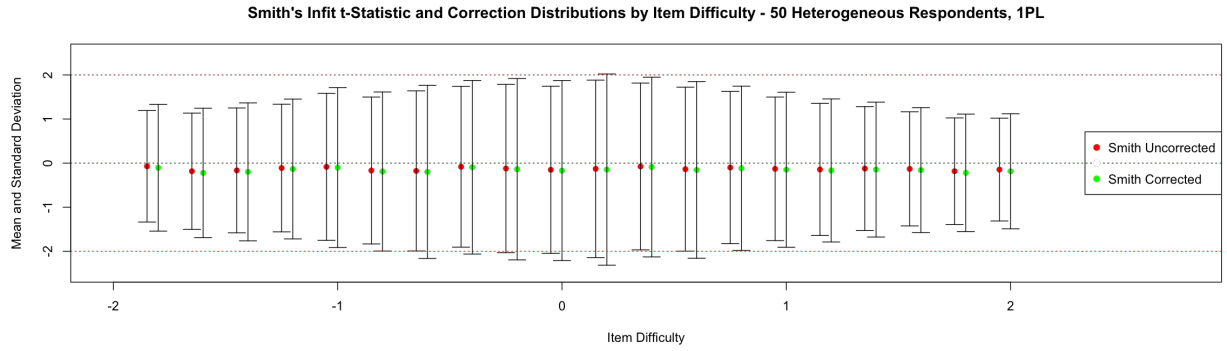


Figure 10.*

**Ignore the Wright and Stone series in this plot. The transformation is not intended for INFIT.*

The Type I errors occur as both positive and negative misfit in the Smith transformation, particularly in the center of the item difficulty distribution, but negative misfit occurs more often. This does not change when the statistics are corrected: the pattern remains unchanged and more Type I errors are observed. The Wright and Masters transformation indicates only negative misfit values, and the total number of false positive indications is unacceptable across all items.

In summary, the null distributions and Type I errors of the mean-squared residuals are very difficult to pin down in small samples. Although increasing sample size does seem to have some impact, the main factor in determining them at the item level seems to be the items location with respect to the center of the ability distribution. The choice of transformation also has a significant impact in small samples. As a result there are no universal recommendations that can be made, other than not to rely on any one statistic or critical value in making a determination on the quality of an item or the data generated from it. [LATER? TYPE I ERROR FOR SOME NOT OTHERS…]

Power. Introducing a controlled amount of model deviation into a simulation can help to evaluate the usefulness of the mean-square statistics and their transformations in detecting misfit.

In order to examine these statistics' ability to detect systematic deviation from the average discrimination in an item, which is the type of model misfit these statistics address (Wu and Adams, In Press), I specify a model that includes six items with a generating discrimination parameter, α_i, of 0.5 to simulate items with an unsatisfactory low slope and six items with α_i set to 1.5 to simulate items that are highly discriminating. These items are also specified to have an equal range of negative and positive difficulty parameters. The remaining 8 items are specified to have a slope of 1, which is equal to the average slope across all items. Therefore, these items should not be indicated with either positive or negative misfit, allowing me to track Type I and II errors simultaneously at varying small sample sizes[2].

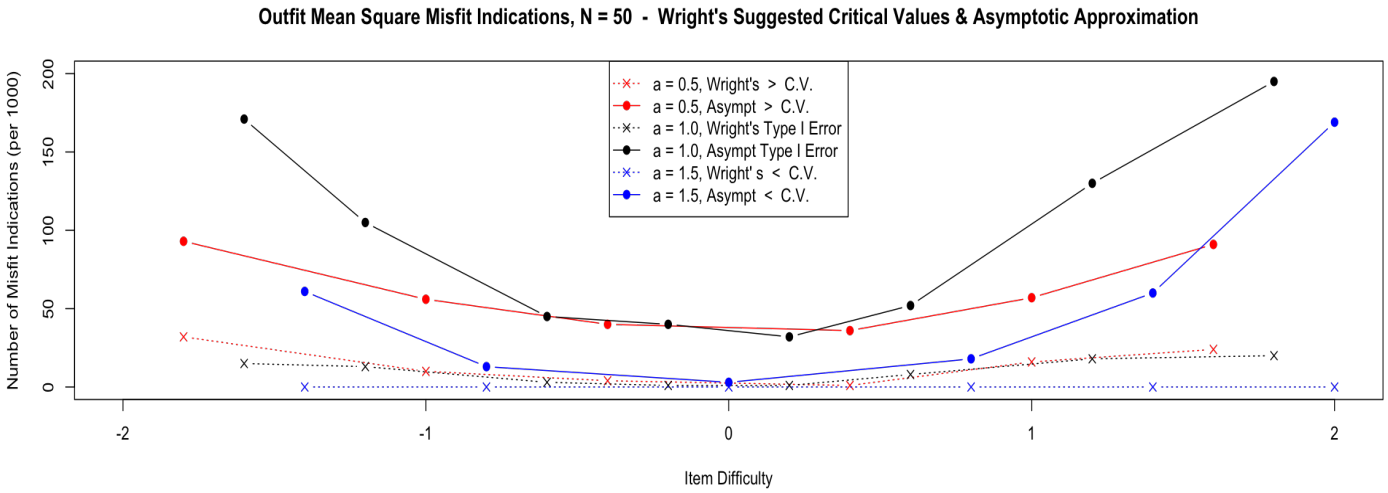
Given this test construction, (which can be interpreted as a restricted 2PL model or three 1PL models nested within one test) my hypothesis is that the six items specified with low discrimination should consistently show positive misfit if the statistics are useful in identifying this type of item. Conversely, six items with a high discrimination parameter should consistently show negative misfit. Therefore, I only consider these expected indications of misfit for the evaluation of power (which reduces the statistical test to a one-tail test). In evaluating Type I error I still consider total misfit indicated for the eight "average" items.

Turning to the OUTFIT mean-square statistic first, the plots of the misfit indications of interest for the three sets of items at a sample size of 50 are shown below in Figure 11. My hypotheses regarding power are not realized: the red and blue series should be elevated, and the black series near zero. The converse has occurred and the differential rates of misfit indication across item difficulties remains. These observations are true regardless of whether the asymptotic or Wright's suggested critical values are used, and does not change as the sample size

[2] I also simulated data with 4 items set to α_i =0.25, 2 items at 2.5 and 14 items at 1.0. Results did not change.

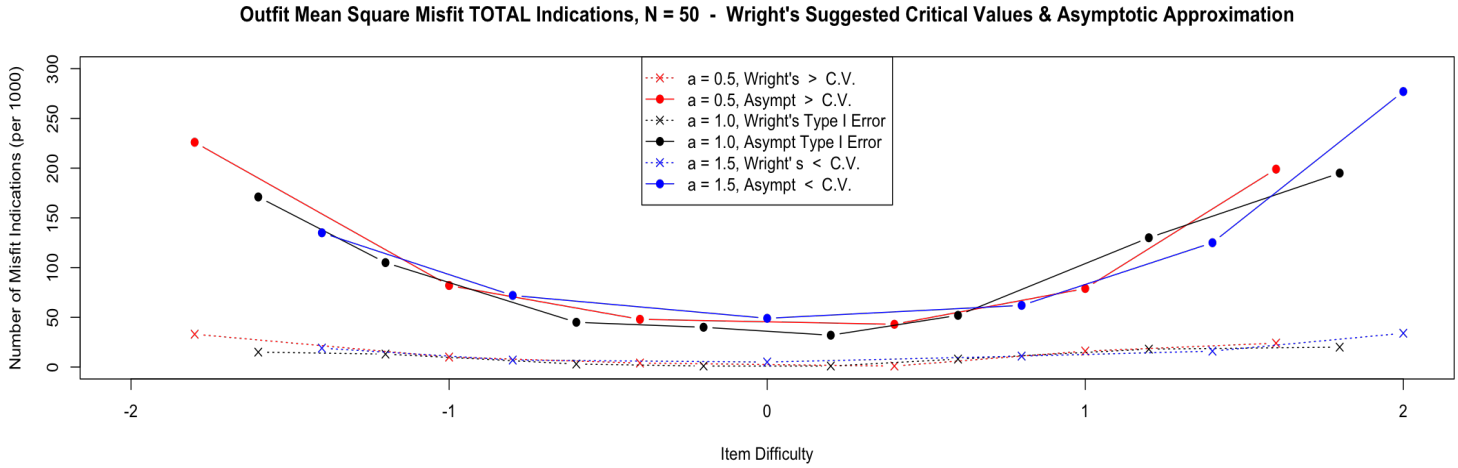
changes from 30 to 70 respondents. This suggests that not only does the OUTFIT statistic lack the necessary power to indicate inconsistent item discrimination, but that it can actually misinform us about items with consistent discrimination.

Figure 11.



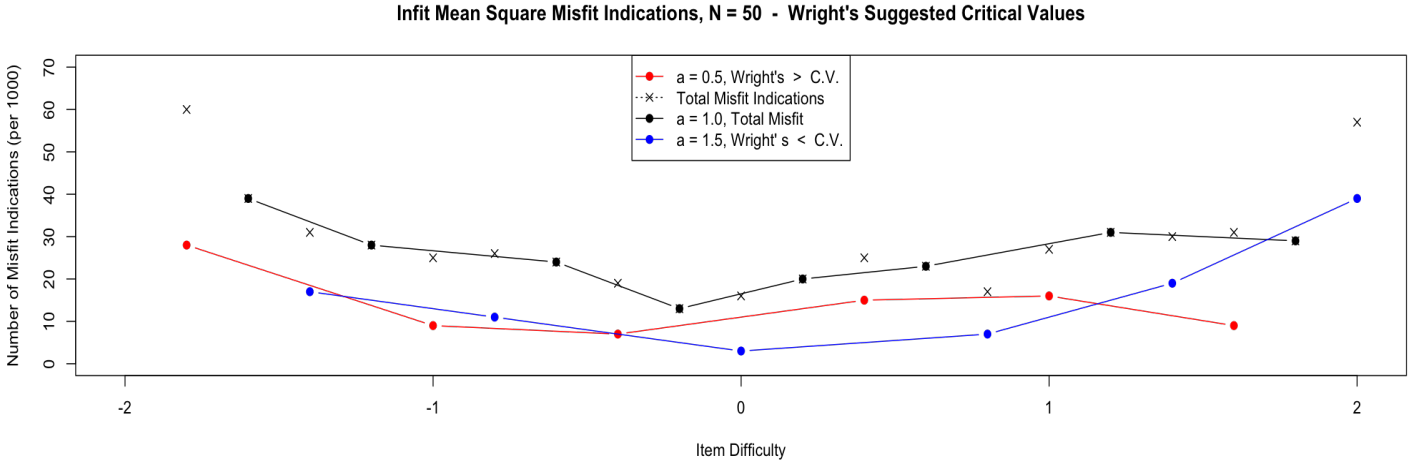
Beyond this problem, the total amount of misfit is not substantially different than what is indicated when the data that is simulated to fit the model. The total here is shown in Figure 12 and can be compared to the sample of 50 from Figure 3 above. This suggests that many of these indications are actually the *opposite* of what we would expect and want to see. This is again due in part to the fact that it is impossible to have negative misfit with Wright's suggested critical values in small samples, but the proportion of observations of positive and negative misfit does not change here either.

Figure 12.



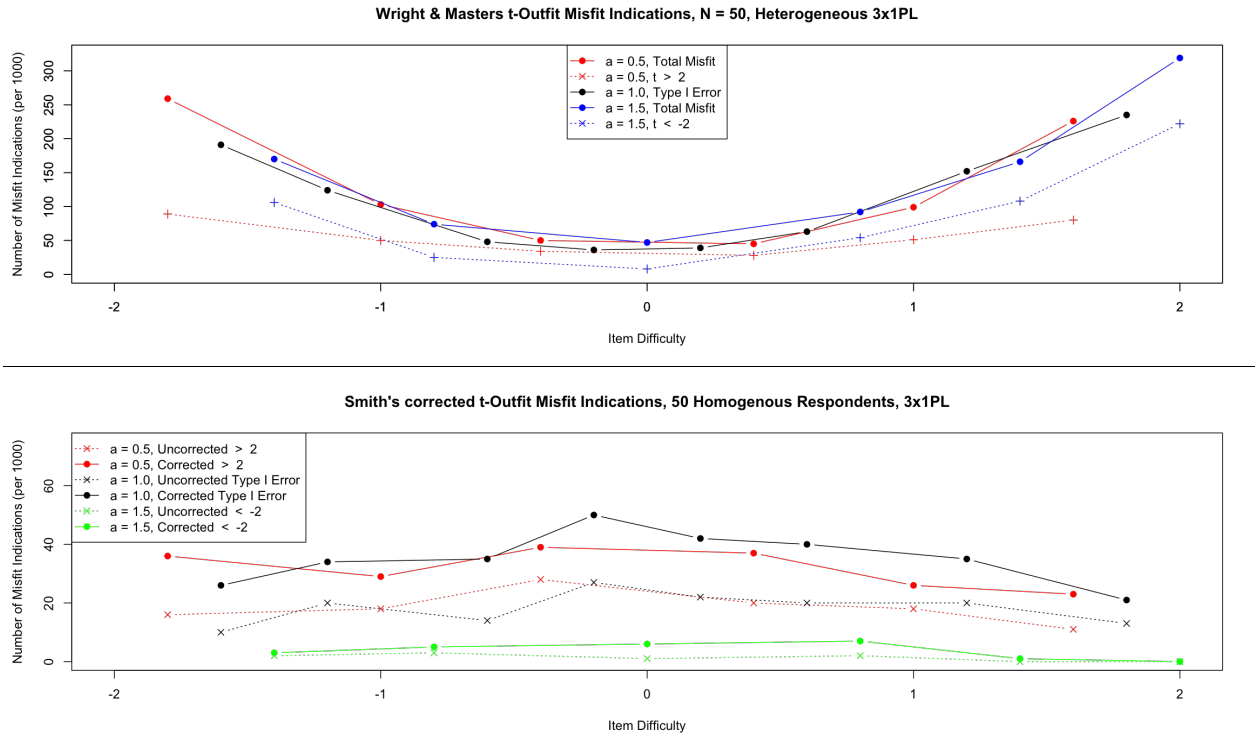
The results for the INFIT mean-square statistic are similar, as shown in Figure 13. The overall rate of misfit is nearly identical to the 1PL simulated data and the items that should be indicated as misfitting more often than not have a lower rate of correct indication than the items specified to fit the model. This is true across all item difficulties. The number of indications of misfit that are in the opposite direction than would be expected for the problematic items is again high (nearly equal to the number of correct indications, as can be seen by looking at the difference between the red and blue dots and the black x's above them in Figure 13). Similar to the OUTFIT statistic, INFIT seems to be more misleading than informative in small samples.

Figure 13.



The standardized OUTFIT and INFIT statistics do not fare any better detecting inconsistent item discrimination. Their distributions are nearly identical to those presented when the model fits the data, and the patterns of expected positive and negative misfit indications for low and highly discriminating items respectively are equally disappointing and troublesome as they are in the raw mean-square statistics. The top panel of Figure 14 demonstrates how the pattern of total misfit indications from the Wright and Masters (1982) transformation, shown in solid dots and lines, continues to be influenced more by item difficulty. The expected misfit indications (shown as "+" signs connected by dotted lines) hardly suggest that these systematically identify misfitting items.

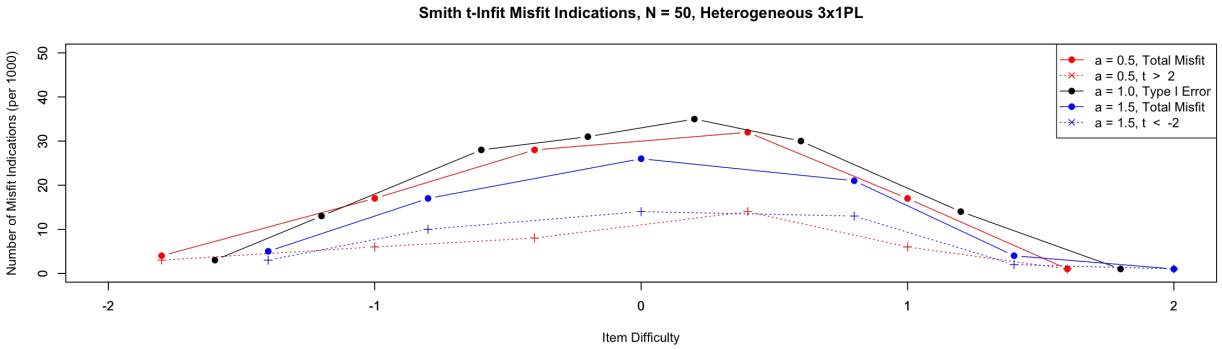
Figure 14.



The second panel of Figure 14 shows that Smith's suggested transformation (dotted lines) and its correction (solid lines) do not perform any better. The indications of misfit continue to follow the distribution of the statistic in relation to each item's distance from the center of ability

distribution. The items that should be shown to fit the model are indicated at an acceptable Type I error rate (50/1000, or 5%), but neither the transformed nor corrected statistic has any power to correctly indicate the misfit specified in the simulation. The same patterns of misfit indications appear in the standardized and corrected INFIT statistics (see Figure 15). The items that are specified to fit the model actually have the highest rate of misfit indication.

Figure 15.

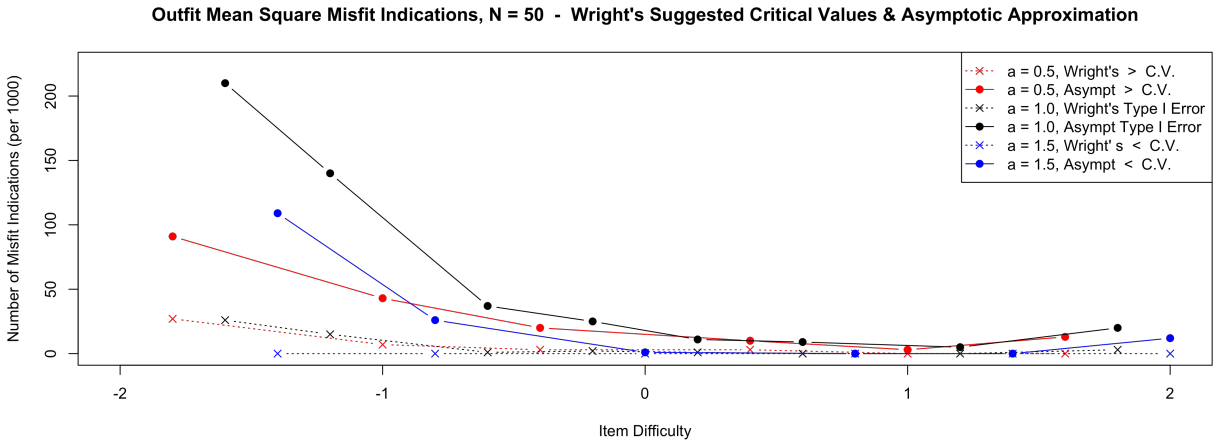


Homogenous ability distributions. Clearly from what I have presented above the indications of item misfit are heavily, if not entirely, influenced by the distance of item difficulty with respect to the location of the center of the ability distribution: items further from the center are indicated more often than those near the center, regardless of the generating item discrimination parameter specified in the simulation. If this is in fact the case, I should expect that manipulating the ability distribution would also change the frequency and pattern of misfit along the continuum of item difficulty. In order to test this hypothesis, I specify an ability distribution that is centered at +0.75 logits and has a standard deviation of 0.5. This is equivalent to a small sample pilot study that uses a high ability group to test the items.

This procedure is conducted on tests simulated with both constant and inconsistent discrimination parameters, but the results of only the latter are presented here. The results from the tests with constant item discriminations are nearly identical. Figure 16 shows the indications

of misfit from the mean-square OUTFIT statistics at a sample size of 50. Here the typical convex shape has been largely flattened out to zero on the upper end of the item difficulty continuum, although it does begin to rise again at the far extreme as this becomes further away from the center of the ability distribution.

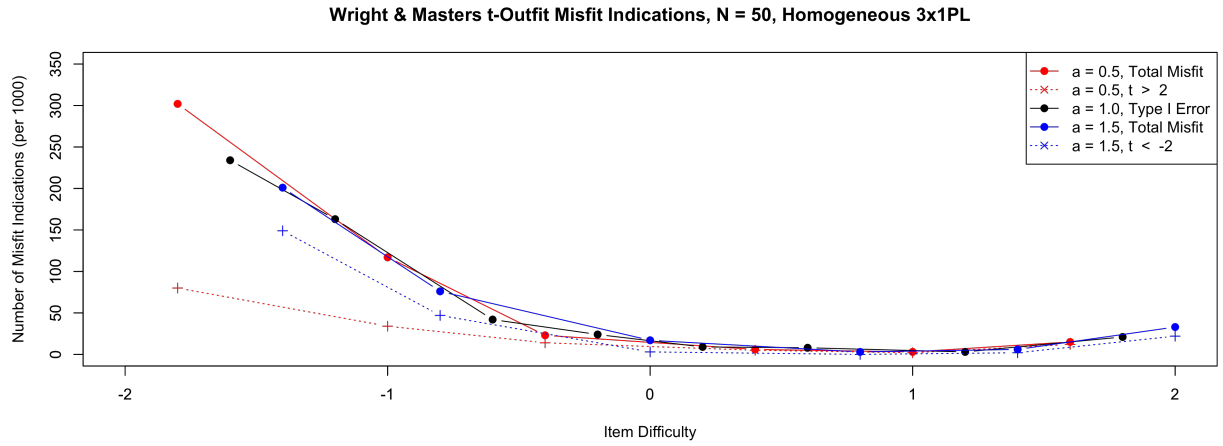
Figure 16.

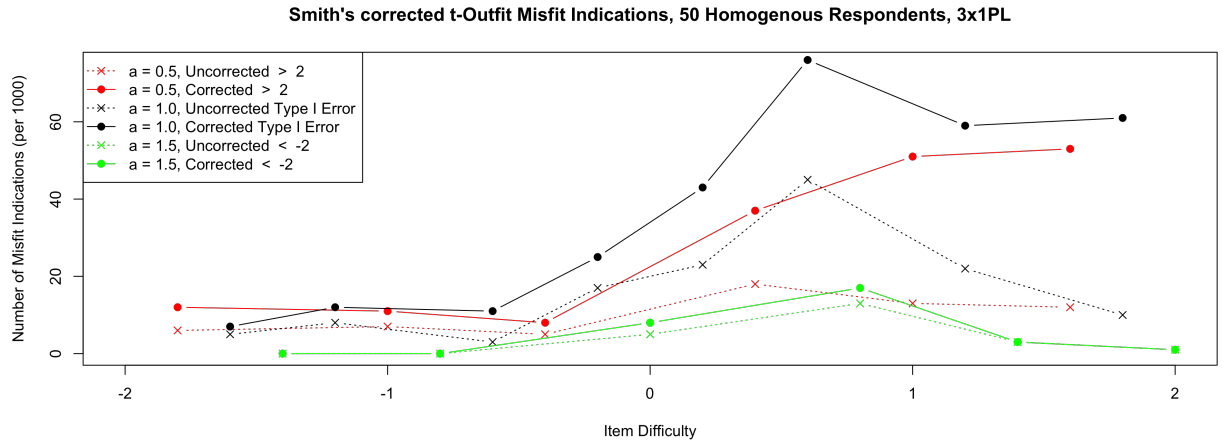


The mean-square INFIT statistic has more indications over the center of the distribution, as was observed in the other conditions. There are also a few indications at the low end of the item difficulty distribution as well, but there are no more than 15 indications per 1000 samples on any one item for any sample size.

Figure 17 shows the plots from the homogenous ability distribution that correspond with the two panels from Figure 14. A similar pattern is observed for the Wright and Masters standardized transformation as the raw mean-square OUTFIT. The general patterns for Smith's transformation and correction are similar to what is observed when the ability distribution is more heterogeneous and disperse, but more extreme. That is, indications of misfit are centered in a way that follows the ability distribution, rather than any systematic relation to the items that actually do misfit the model.

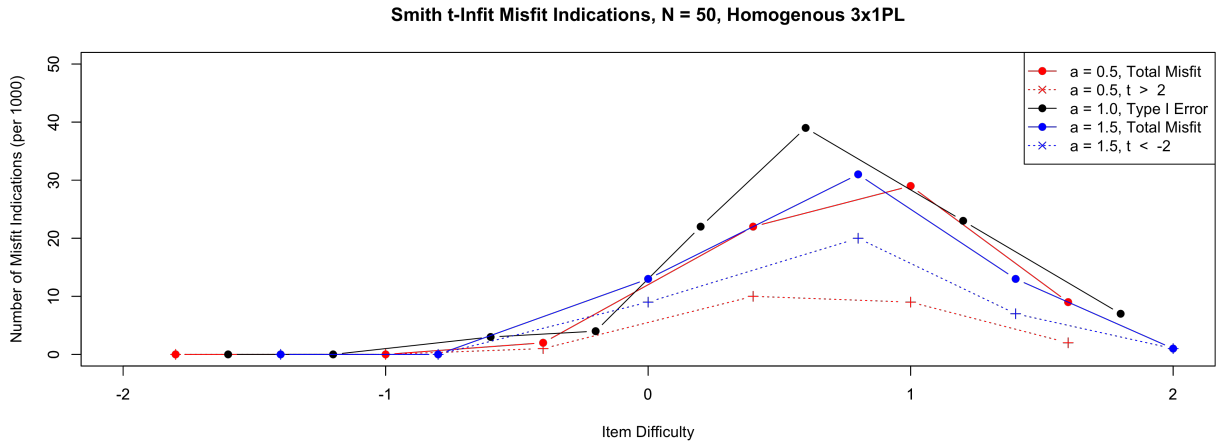
Figure 17.

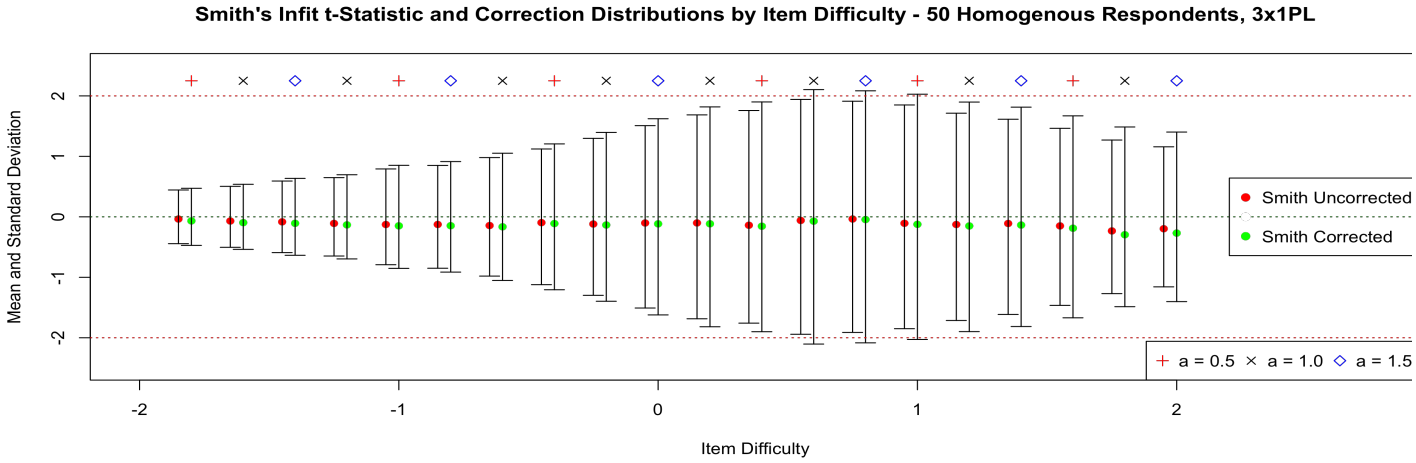




Finally, the pattern of misfit from Smith's transformation and the distribution of it and the proposed correction are shown in Figure 18. Misfit continues to be indicated based on the ability distribution, and the effect that has on the item specific distribution of the statistic rather than dependent upon actual misfit of the data.

Figure 18.


Smith t-Infit Misfit Indications, N = 50, Homogenous 3x1PL
Number of Misfit Indications (per 1000)
0
10
20
30
40
50
-2
-1
0
1
2
Item Difficulty
a = 0.5, Total Misfit
a = 0.5, t > 2
a = 1.0, Type I Error
a = 1.5, Total Misfit
a = 1.5, t < -2


Smith's Infit t-Statistic and Correction Distributions by Item Difficulty - 50 Homogenous Respondents, 3x1PL
Mean and Standard Deviation
-2
-1
0
1
2
-2
-1
0
1
2
Item Difficulty
Smith Uncorrected
Smith Corrected
a = 0.5 a = 1.0 a = 1.5

Discussion

Things to do: References, formulae, tables and figures, etc.

Adams, R. and Khoo, S. (1996). *ACER Quest – The Interactive Test Analysis System.* ACER Press, Camberwell.

George, A. (1979). Theoretical and practical consequences of the use of standardised residuals as Rasch model fit statistics. *Paper presented at the 1979 Annual Meeting of the American Educational Research Association*.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, 1(2), 152-176.

Fischer, G., and Molenaar, I. (1995) (Eds.). *Rasch Models –Foundations, Recent Developments, and Applications.* NY: Springer-Verlag.

R Development Core Team (2009). R: A language and environment for statistical computing. Vienna, Austria: Author (3- 900051-07-0).

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy,* 14, 58-94.

Smith, R. (1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measurement,* 48, 657-667.

Smith, R. (1990). Theory and practice of fit. *Rasch Measurement Transactions (RMT),* 3:4, p.78 .

Smith, R. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement,* 51, 541-565.

Smith, R., Schumacker, R., & Bush, M. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2(1), 66-78.

Wang, W., and Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement*, 65(3), 376-404.

Wilson, E., and Hilferty, M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America,* 17, 684-688.

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

Wright, B., and Linacre, J. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions,* 1994, 8:3 p.370 .

Wright, B., and Masters, G. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B., and Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-48.

APPENDIX. SELECTED TABLES USED TO CONSTRUCT PLOTS.

Table A0. Means and Standard Deviations of Mean-square OUTFIT Statistics

		N = 30		*N = 50*		*N = 70*	
	Delta	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Item 1	**-1.8**	0.97	*0.20*	0.97	*0.15*	0.97	*0.12*
Item 2	**-1.6**	0.95	*0.18*	0.95	*0.13*	0.96	*0.11*
Item 3	**-1.4**	0.95	*0.18*	0.96	*0.13*	0.96	*0.11*
Item 4	**-1.2**	0.96	*0.18*	0.97	*0.13*	0.97	*0.11*
Item 5	**-1.0**	0.97	*0.16*	0.98	*0.13*	0.97	*0.11*
Item 6	**-0.8**	0.97	*0.17*	0.97	*0.12*	0.97	*0.10*
Item 7	**-0.6**	0.99	*0.17*	0.97	*0.13*	0.98	*0.11*
Item 8	**-0.4**	0.98	*0.16*	0.99	*0.12*	0.98	*0.11*
Item 9	**-0.2**	0.99	*0.16*	0.98	*0.12*	0.99	*0.11*
Item 10	**0.0**	0.98	*0.16*	0.98	*0.12*	0.99	*0.11*
Item 11	**0.2**	0.99	*0.17*	0.98	*0.13*	0.99	*0.10*
Item 12	**0.4**	0.98	*0.16*	0.99	*0.13*	0.98	*0.11*
Item 13	**0.6**	0.98	*0.16*	0.98	*0.13*	0.98	*0.10*
Item 14	**0.8**	0.97	*0.16*	0.98	*0.13*	0.98	*0.11*
Item 15	**1.0**	0.98	*0.17*	0.98	*0.13*	0.98	*0.11*
Item 16	**1.2**	0.96	*0.17*	0.97	*0.13*	0.97	*0.11*
Item 17	**1.4**	0.96	*0.17*	0.97	*0.13*	0.97	*0.11*
Item 18	**1.6**	0.95	*0.18*	0.96	*0.13*	0.96	*0.11*
Item 19	**1.8**	0.95	*0.18*	0.95	*0.13*	0.95	*0.11*
Item 20	**2.0**	0.94	*0.19*	0.95	*0.14*	0.96	*0.12*

Table A1. Type I Error Rates for Outfit Mean Squares (Per 1,000 Samples) and Range of Values

N = 30		Wright's Suggested CV			Asymptotic CV				
	Delta	**Low**	**High**	**Total**	**Low**	**High**	**Total**	**Min 30**	**Max 30**
Item 1	**-1.8**	0	35	35	117	77	194	0.07	14.08
Item 2	**-1.6**	0	22	22	80	73	153	0.05	4.70
Item 3	**-1.4**	0	23	23	60	68	128	0.18	6.75
Item 4	**-1.2**	0	20	20	46	65	111	0.27	5.16
Item 5	**-1.0**	0	20	20	21	68	89	0.25	7.47
Item 6	**-0.8**	0	7	7	13	54	67	0.17	3.35
Item 7	**-0.6**	0	16	16	8	62	70	0.32	2.93
Item 8	**-0.4**	0	7	7	5	44	49	0.36	3.26
Item 9	**-0.2**	0	7	7	5	43	48	0.43	2.92
Item 10	**0.0**	0	8	8	1	37	38	0.46	3.26
Item 11	**0.2**	0	9	9	3	45	48	0.39	4.49
Item 12	**0.4**	0	13	13	5	58	63	0.41	4.80
Item 13	**0.6**	0	5	5	8	40	48	0.34	3.26
Item 14	**0.8**	0	8	8	15	51	66	0.33	3.37
Item 15	**1.0**	0	13	13	19	61	80	0.24	4.09
Item 16	**1.2**	0	18	18	35	66	101	0.13	3.89
Item 17	**1.4**	0	20	20	57	71	128	0.18	7.56
Item 18	**1.6**	0	22	22	74	71	145	0.09	7.68
Item 19	**1.8**	0	24	24	112	90	202	0.13	11.96
Item 20	**2.0**	0	34	34	153	82	235	0.08	12.60
				0.0165	total proportion		0.103	total proportion	
		0	331	331.0	837	1226	2063		

N = 50		Wright's Suggested CV			Asymptotic CV				
	Delta	**Low**	**High**	**Total**	**Low**	**High**	**Total**	**Min 50**	**Max 50**
Item 1	**-1.8**	0	20	20	130	96	226	0.22	3.07
Item 2	**-1.6**	0	19	19	88	68	156	0.29	3.56
Item 3	**-1.4**	0	15	15	86	65	151	0.32	4.8
Item 4	**-1.2**	0	18	18	44	74	118	0.32	2.62
Item 5	**-1.0**	0	10	10	24	55	79	0.45	2.84
Item 6	**-0.8**	0	8	8	14	49	63	0.51	2.99
Item 7	**-0.6**	0	1	1	16	41	57	0.48	1.98
Item 8	**-0.4**	0	7	7	5	42	47	0.56	2.51
Item 9	**-0.2**	0	3	3	5	37	42	0.58	2.25
Item 10	**0.0**	0	3	3	6	30	36	0.51	2.15
Item 11	**0.2**	0	5	5	5	49	54	0.5	2.67
Item 12	**0.4**	0	5	5	5	39	44	0.42	2.83
Item 13	**0.6**	0	6	6	23	47	70	0.44	2.77

Item 14	**0.8**	0	12	12		12	57	69	0.51	2.83
Item 15	**1.0**	0	14	14		27	59	86	0.39	2.93
Item 16	**1.2**	0	11	11		36	60	96	0.47	3.42
Item 17	**1.4**	0	18	18		68	64	132	0.3	4.16
Item 18	**1.6**	0	25	25		87	87	174	0.2	4.24
Item 19	**1.8**	0	24	24		131	83	214	0.24	6.57
Item 20	**2.0**	0	30	30		158	78	236	0.18	4.28
				0.0127		total proportion		0.108	total proportion	
		0	254	254		970	1180	2150		

N = 70 Wright's Suggested CV Asymptotic CV

	Delta	**Low**	**High**	**Total**		**Low**	**High**	**Total**		**Min 70**	**Max 70**
Item 1	**-1.8**	0	30	30		122	105	227		0.33	3.54
Item 2	**-1.6**	1	21	22		89	78	167		0.23	3.95
Item 3	**-1.4**	0	13	13		89	72	161		0.42	2.93
Item 4	**-1.2**	0	8	8		59	52	111		0.48	2.79
Item 5	**-1.0**	0	12	12		47	57	104		0.55	3.59
Item 6	**-0.8**	0	5	5		28	32	60		0.56	2.56
Item 7	**-0.6**	0	8	8		16	38	54		0.56	2.22
Item 8	**-0.4**	0	3	3		12	26	38		0.57	2.12
Item 9	**-0.2**	0	5	5		4	47	51		0.6	2.25
Item 10	**0.0**	0	7	7		6	45	51		0.56	1.91
Item 11	**0.2**	0	4	4		7	41	48		0.61	2.11
Item 12	**0.4**	0	7	7		16	38	54		0.55	2.21
Item 13	**0.6**	0	8	8		14	41	55		0.58	2.29
Item 14	**0.8**	0	8	8		31	45	76		0.48	2.64
Item 15	**1.0**	0	11	11		51	62	113		0.45	2.79
Item 16	**1.2**	0	12	12		55	75	130		0.49	2.32
Item 17	**1.4**	0	11	11		87	60	147		0.48	2.71
Item 18	**1.6**	0	22	22		122	81	203		0.42	4.2
Item 19	**1.8**	0	35	35		158	94	252		0.31	4.48
Item 20	**2.0**	2	28	30		173	106	279		0.23	6.23
				0.0130 5		total proportion		0.119		total proportion	
		3	258	261		1186	1195	2381			

Table A2. Type I Error Rates for Infit Mean Squares, and Minimum and Maximum Values

		N = 30			*Infit Values*		*N = 50*			*Infit Values*		*N = 70*			*Infit Values*	
	Delta	**Low**	**High**	**Total**	**Min30**	**Max30**	**Low**	**High**	**Total**	**Min50**	**Max50**	**Low**	**High**	**Total**	**Min70**	**Max70**
Item 1	**-1.8**	39	27	66	0.35	1.65	37	21	58	0.56	1.64	31	13	44	0.66	1.39
Item 2	**-1.6**	23	11	34	0.26	1.59	30	11	41	0.56	1.37	35	5	40	0.64	1.36
Item 3	**-1.4**	23	13	36	0.42	1.53	27	15	42	0.57	1.44	31	7	38	0.61	1.31
Item 4	**-1.2**	25	22	47	0.5	1.60	27	9	36	0.61	1.40	29	8	37	0.67	1.37
Item 5	**-1.0**	14	13	27	0.43	1.5	12	13	25	0.66	1.39	18	10	28	0.68	1.33
Item 6	**-0.8**	10	17	27	0.42	1.77	7	10	17	0.67	1.55	19	8	27	0.68	1.40
Item 7	**-0.6**	12	23	35	0.52	1.71	19	9	28	0.60	1.38	15	13	28	0.65	1.42
Item 8	**-0.4**	8	11	19	0.48	1.51	11	10	21	0.67	1.48	12	14	26	0.69	1.30
Item 9	**-0.2**	9	18	27	0.55	1.71	7	8	15	0.66	1.37	8	20	28	0.71	1.45
Item 10	**0.0**	10	17	27	0.59	1.67	12	5	17	0.61	1.38	10	14	24	0.69	1.46
Item 11	**0.2**	8	29	37	0.52	1.86	10	20	30	0.62	1.44	6	11	17	0.73	1.37
Item 12	**0.4**	11	13	24	0.5	1.52	7	15	22	0.57	1.46	13	10	23	0.68	1.38
Item 13	**0.6**	6	11	17	0.51	1.79	20	14	34	0.59	1.43	11	7	18	0.71	1.43
Item 14	**0.8**	11	11	22	0.55	1.52	9	14	23	0.67	1.55	22	9	31	0.59	1.31
Item 15	**1.0**	15	17	32	0.49	1.76	14	8	22	0.62	1.42	19	11	30	0.66	1.31
Item 16	**1.2**	12	14	26	0.47	1.90	13	10	23	0.66	1.44	18	13	31	0.68	1.39
Item 17	**1.4**	16	15	31	0.42	1.67	20	11	31	0.58	1.59	27	10	37	0.7	1.31
Item 18	**1.6**	24	15	39	0.44	1.67	24	17	41	0.59	1.36	21	12	33	0.67	1.40
Item 19	**1.8**	30	16	46	0.42	1.57	32	7	39	0.54	1.46	47	7	54	0.6	1.31
Item 20	**2.0**	38	23	61	0.36	1.71	38	18	56	0.49	1.37	43	14	57	0.61	1.37
		0.017	0.017	0.034			0.019	0.012	0.031			0.022	0.011	0.033		

Table A3. Means and Standard Deviations of Three *t*-Outfit Statistics

N = 30		***Wright & Stone***		***Smith***		***Wright & Masters***		***Aysmptotic Approximation***	
	Delta	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Item 1	**-1.8**	-0.47	*2.23*	0.02	*0.72*	-2.18	*4.96*	-0.48	*1.95*
Item 2	**-1.6**	-0.50	*1.57*	-0.05	*0.69*	-2.32	*6.45*	-0.47	*1.55*
Item 3	**-1.4**	-0.34	*1.72*	-0.02	*0.78*	-2.10	*6.19*	-0.32	*1.61*
Item 4	**-1.2**	-0.32	*1.47*	-0.04	*0.79*	-2.20	*9.24*	-0.28	*1.43*
Item 5	**-1.0**	-0.19	*1.51*	0.00	*0.85*	-1.50	*3.60*	-0.15	*1.40*
Item 6	**-0.8**	-0.25	*1.17*	-0.06	*0.82*	-1.48	*3.11*	-0.20	*1.17*
Item 7	**-0.6**	-0.16	*1.17*	-0.03	*0.89*	-1.53	*3.55*	-0.10	*1.16*
Item 8	**-0.4**	-0.16	*1.08*	-0.03	*0.87*	-1.55	*3.56*	-0.10	*1.07*
Item 9	**-0.2**	-0.15	*1.02*	-0.04	*0.88*	-1.65	*4.48*	-0.09	*1.02*
Item 10	**0.0**	-0.18	*1.05*	-0.08	*0.91*	-1.56	*4.12*	-0.13	*1.04*
Item 11	**0.2**	-0.11	*1.09*	-0.02	*0.90*	-1.25	*2.85*	-0.05	*1.06*
Item 12	**0.4**	-0.12	*1.19*	-0.02	*0.92*	-1.37	*3.73*	-0.07	*1.16*
Item 13	**0.6**	-0.21	*1.02*	-0.06	*0.78*	-1.48	*3.87*	-0.15	*1.02*
Item 14	**0.8**	-0.25	*1.21*	-0.05	*0.83*	-1.84	*4.97*	-0.20	*1.20*
Item 15	**1.0**	-0.22	*1.26*	-0.02	*0.78*	-1.50	*4.01*	-0.18	*1.25*
Item 16	**1.2**	-0.33	*1.36*	-0.04	*0.76*	-1.70	*5.32*	-0.29	*1.34*
Item 17	**1.4**	-0.38	*1.68*	-0.03	*0.72*	-3.37	*42.22*	-0.36	*1.56*
Item 18	**1.6**	-0.45	*1.77*	-0.03	*0.68*	-1.85	*4.01*	-0.44	*1.65*
Item 19	**1.8**	-0.50	*2.02*	0.03	*0.69*	-2.17	*5.46*	-0.50	*1.86*
Item 20	**2.0**	-0.64	*2.25*	0.03	*0.67*	-2.39	*5.06*	-0.66	*2.06*

N = 50		***Wright & Stone***		***Smith***		***Wright & Masters***		***Aysmptotic Approx***	
	Delta	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Item 1	**-1.8**	-0.48	*1.74*	-0.04	*0.69*	-1.35	*3.34*	-0.47	*1.73*
Item 2	**-1.6**	-0.50	*1.54*	-0.11	*0.72*	-1.41	*3.14*	-0.48	*1.53*
Item 3	**-1.4**	-0.43	*1.56*	-0.10	*0.76*	-1.54	*3.90*	-0.41	*1.52*
Item 4	**-1.2**	-0.26	*1.36*	-0.04	*0.76*	-1.15	*3.31*	-0.22	*1.35*
Item 5	**-1.0**	-0.23	*1.25*	-0.06	*0.81*	-1.31	*3.23*	-0.19	*1.24*
Item 6	**-0.8**	-0.27	*1.15*	-0.11	*0.83*	-1.59	*6.01*	-0.23	*1.15*
Item 7	**-0.6**	-0.25	*1.06*	-0.12	*0.85*	-1.41	*3.79*	-0.21	*1.06*
Item 8	**-0.4**	-0.13	*1.06*	-0.05	*0.89*	-1.26	*4.65*	-0.08	*1.05*
Item 9	**-0.2**	-0.18	*0.98*	-0.11	*0.89*	-1.27	*2.99*	-0.14	*0.99*
Item 10	**0.0**	-0.19	*1.00*	-0.11	*0.92*	-1.86	*9.73*	-0.14	*1.00*
Item 11	**0.2**	-0.18	*1.06*	-0.09	*0.95*	-1.54	*4.86*	-0.13	*1.06*
Item 12	**0.4**	-0.13	*1.06*	-0.05	*0.89*	-1.21	*3.28*	-0.09	*1.06*
Item 13	**0.6**	-0.21	*1.11*	-0.10	*0.89*	-1.43	*4.52*	-0.17	*1.11*
Item 14	**0.8**	-0.22	*1.20*	-0.08	*0.86*	-1.34	*3.18*	-0.18	*1.18*
Item 15	**1.0**	-0.25	*1.29*	-0.07	*0.84*	-1.45	*4.45*	-0.21	*1.27*

Item 16	**1.2**	-0.33	*1.35*	-0.09	*0.80*	-1.47	*3.97*	-0.29	*1.34*
Item 17	**1.4**	-0.35	*1.50*	-0.07	*0.73*	-1.31	*3.38*	-0.33	*1.47*
Item 18	**1.6**	-0.40	*1.68*	-0.06	*0.75*	-1.93	*15.16*	-0.38	*1.64*
Item 19	**1.8**	-0.55	*1.86*	-0.08	*0.72*	-1.68	*4.61*	-0.54	*1.78*
Item 20	**2.0**	-0.59	*1.93*	-0.06	*0.68*	-1.55	*3.93*	-0.59	*1.89*

N = 70		***Wright & Stone***		***Smith***		***Wright & Masters***		***Aysmptotic Approximation***	
	Delta	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Item 1	**-1.8**	-0.36	*1.85*	-0.03	*0.73*	-0.89	*2.14*	-0.36	*1.81*
Item 2	**-1.6**	-0.49	*1.64*	-0.13	*0.72*	-1.15	*3.05*	-0.48	*1.62*
Item 3	**-1.4**	-0.43	*1.50*	-0.13	*0.77*	-1.21	*3.47*	-0.41	*1.49*
Item 4	**-1.2**	-0.37	*1.34*	-0.14	*0.79*	-1.65	*9.42*	-0.35	*1.33*
Item 5	**-1.0**	-0.33	*1.31*	-0.14	*0.84*	-1.57	*6.90*	-0.30	*1.29*
Item 6	**-0.8**	-0.33	*1.17*	-0.18	*0.84*	-1.20	*2.81*	-0.30	*1.16*
Item 7	**-0.6**	-0.25	*1.11*	-0.14	*0.89*	-1.36	*5.37*	-0.22	*1.11*
Item 8	**-0.4**	-0.19	*1.03*	-0.11	*0.90*	-1.34	*4.60*	-0.15	*1.03*
Item 9	**-0.2**	-0.10	*1.07*	-0.04	*0.96*	-1.21	*6.90*	-0.06	*1.06*
Item 10	**0.0**	-0.09	*1.04*	-0.04	*0.94*	-1.06	*3.36*	-0.05	*1.03*
Item 11	**0.2**	-0.16	*1.04*	-0.10	*0.93*	-1.17	*3.72*	-0.12	*1.03*
Item 12	**0.4**	-0.24	*1.09*	-0.15	*0.94*	-1.36	*4.13*	-0.20	*1.08*
Item 13	**0.6**	-0.27	*1.10*	-0.17	*0.85*	-1.13	*2.76*	-0.24	*1.09*
Item 14	**0.8**	-0.24	*1.21*	-0.11	*0.88*	-1.32	*5.35*	-0.21	*1.20*
Item 15	**1.0**	-0.28	*1.35*	-0.11	*0.85*	-1.31	*4.04*	-0.26	*1.33*
Item 16	**1.2**	-0.31	*1.41*	-0.09	*0.82*	-1.13	*2.94*	-0.28	*1.40*
Item 17	**1.4**	-0.44	*1.42*	-0.13	*0.74*	-1.14	*2.81*	-0.41	*1.42*
Item 18	**1.6**	-0.47	*1.70*	-0.11	*0.75*	-1.36	*5.03*	-0.46	*1.66*
Item 19	**1.8**	-0.54	*1.99*	-0.11	*0.77*	-1.45	*3.98*	-0.54	*1.92*
Item 20	**2.0**	-0.49	*2.14*	-0.04	*0.73*	-1.28	*3.75*	-0.50	*2.05*

Table A4. Means and Standard Deviations of Three *t*-Infit Statistics

N = 30		***Wright & Stone***		***Smith***		***Wright & Masters***	
	Delta	Mean	SD	Mean	SD	Mean	SD
Item 1	**-1.8**	-0.18	0.78	-0.04	0.66	-0.39	1.41
Item 2	**-1.6**	-0.22	0.70	-0.11	0.68	-0.57	1.83
Item 3	**-1.4**	-0.21	0.69	-0.13	0.73	-0.63	1.97
Item 4	**-1.2**	-0.18	0.69	-0.12	0.80	-0.81	3.52
Item 5	**-1.0**	-0.14	0.63	-0.11	0.78	-0.66	2.06
Item 6	**-0.8**	-0.13	0.66	-0.11	0.87	-0.88	3.90
Item 7	**-0.6**	-0.08	0.65	-0.07	0.93	-1.00	9.43
Item 8	**-0.4**	-0.09	0.62	-0.08	0.91	-0.86	3.42
Item 9	**-0.2**	-0.07	0.62	-0.06	0.94	-0.71	2.10
Item 10	**0.0**	-0.11	0.63	-0.11	0.97	-0.90	2.59
Item 11	**0.2**	-0.05	0.63	-0.04	0.95	-0.85	4.80
Item 12	**0.4**	-0.09	0.61	-0.08	0.92	-0.72	2.32
Item 13	**0.6**	-0.09	0.60	-0.07	0.85	-0.75	3.09
Item 14	**0.8**	-0.14	0.64	-0.13	0.85	-0.90	3.08
Item 15	**1.0**	-0.10	0.65	-0.05	0.83	-0.62	2.19
Item 16	**1.2**	-0.17	0.65	-0.11	0.77	-0.60	1.89
Item 17	**1.4**	-0.20	0.66	-0.12	0.70	-0.57	2.53
Item 18	**1.6**	-0.22	0.69	-0.11	0.67	-0.58	2.05
Item 19	**1.8**	-0.23	0.71	-0.08	0.61	-0.44	2.05
Item 20	**2.0**	-0.26	0.75	-0.08	0.58	-0.43	1.54

N = 50		***Wright & Stone***		***Smith***		***Wright & Masters***	
	Delta	Mean	SD	Mean	SD	Mean	SD
Item 1	**-1.8**	-0.18	0.74	-0.07	0.63	-0.26	1.00
Item 2	**-1.6**	-0.27	0.67	-0.18	0.66	-0.44	1.19
Item 3	**-1.4**	-0.23	0.67	-0.17	0.71	-0.51	2.13
Item 4	**-1.2**	-0.15	0.63	-0.11	0.72	-0.51	3.79
Item 5	**-1.0**	-0.10	0.65	-0.08	0.83	-0.68	6.79
Item 6	**-0.8**	-0.16	0.62	-0.17	0.83	-0.66	2.30
Item 7	**-0.6**	-0.16	0.63	-0.18	0.91	-1.00	8.45
Item 8	**-0.4**	-0.08	0.61	-0.08	0.91	-0.59	2.10
Item 9	**-0.2**	-0.10	0.61	-0.12	0.95	-0.83	3.29
Item 10	**0.0**	-0.12	0.60	-0.15	0.95	-0.72	2.29
Item 11	**0.2**	-0.11	0.65	-0.13	1.01	-0.81	3.17
Item 12	**0.4**	-0.08	0.62	-0.08	0.95	-0.66	3.94
Item 13	**0.6**	-0.12	0.64	-0.14	0.93	-0.74	2.62
Item 14	**0.8**	-0.11	0.63	-0.10	0.86	-0.54	1.86
Item 15	**1.0**	-0.14	0.63	-0.13	0.81	-0.58	2.55
Item 16	**1.2**	-0.18	0.63	-0.14	0.75	-0.46	1.61

Item 17	**1.4**	-0.18	0.66	-0.12	0.70	-0.57	6.20
Item 18	**1.6**	-0.21	0.65	-0.13	0.65	-0.35	1.20
Item 19	**1.8**	-0.30	0.68	-0.18	0.60	-0.39	1.07
Item 20	**2.0**	-0.30	0.72	-0.15	0.58	-0.34	1.02

N = 70		***Wright & Stone***		***Smith***		***Wright & Masters***	
	Delta	Mean	SD	Mean	SD	Mean	SD
Item 1	**-1.8**	-0.19	0.72	-0.10	0.63	-0.23	0.78
Item 2	**-1.6**	-0.28	0.65	-0.21	0.62	-0.37	0.86
Item 3	**-1.4**	-0.25	0.66	-0.20	0.71	-0.42	1.14
Item 4	**-1.2**	-0.21	0.65	-0.20	0.77	-0.47	1.34
Item 5	**-1.0**	-0.19	0.64	-0.19	0.82	-0.53	1.78
Item 6	**-0.8**	-0.18	0.62	-0.21	0.85	-0.59	1.83
Item 7	**-0.6**	-0.14	0.64	-0.17	0.93	-0.56	1.71
Item 8	**-0.4**	-0.11	0.63	-0.14	0.96	-1.03	14.08
Item 9	**-0.2**	-0.06	0.63	-0.06	0.98	-0.49	2.14
Item 10	**0.0**	-0.06	0.62	-0.06	0.97	-0.49	2.10
Item 11	**0.2**	-0.10	0.61	-0.13	0.95	-0.57	2.06
Item 12	**0.4**	-0.13	0.64	-0.17	0.98	-0.79	3.45
Item 13	**0.6**	-0.15	0.60	-0.18	0.87	-0.54	1.65
Item 14	**0.8**	-0.16	0.64	-0.18	0.89	-0.74	4.01
Item 15	**1.0**	-0.17	0.65	-0.17	0.82	-0.46	1.47
Item 16	**1.2**	-0.18	0.65	-0.16	0.76	-0.38	1.12
Item 17	**1.4**	-0.21	0.64	-0.17	0.69	-0.36	1.02
Item 18	**1.6**	-0.27	0.66	-0.20	0.65	-0.37	0.89
Item 19	**1.8**	-0.32	0.68	-0.21	0.61	-0.39	1.12
Item 20	**2.0**	-0.25	0.73	-0.12	0.58	-0.25	0.80

Table A5. Type I Error Rates for Outfit Mean Squares (Per 1,000 Samples)
Data simulated with multiple discrimination values (Alpha)

N = 30

	Delta	Alpha	Low	High	Total	Low	High	Total	Min 30	Max 30
Item 1	**-1.8**	**0.5**	0	42	42	5	172	177	0.38	4.11
Item 2	**-1.0**	**0.5**	0	30	30	0	145	145	0.51	7.07
Item 3	**-0.4**	**0.5**	0	19	19	0	130	130	0.53	4.58
Item 4	**0.4**	**0.5**	0	16	16	0	159	159	0.49	4.33
Item 5	**1.0**	**0.5**	0	29	29	0	150	150	0.5	3.81
Item 6	**1.6**	**0.5**	0	44	44	3	173	176	0.39	4.62
Item 7	**-1.6**	**1**	0	22	22	76	62	138	0.15	5.61
Item 8	**-1.2**	**1**	0	5	5	31	44	75	0.16	4.57
Item 9	**-0.6**	**1**	0	4	4	5	19	24	0.33	3.83
Item 10	**-0.2**	**1**	0	5	5	2	23	25	0.42	3.84
Item 11	**0.2**	**1**	0	3	3	1	26	27	0.44	2.25
Item 12	**0.6**	**1**	0	3	3	9	39	48	0.36	3.7
Item 13	**1.2**	**1**	0	8	8	31	50	81	0.09	7.98
Item 14	**1.8**	**1**	0	29	29	107	69	176	0.11	6.04
Item 15	**-1.4**	**1.5**	0	10	10	201	23	224	0.13	4.85
Item 16	**-0.8**	**1.5**	0	2	2	40	10	50	0.3	5.19
Item 17	**0.0**	**1.5**	0	3	3	11	7	18	0.35	3.18
Item 18	**0.8**	**1.5**	0	0	0	52	6	58	0.26	1.75
Item 19	**1.4**	**1.5**	0	8	8	200	23	223	0.05	5.61
Item 20	**2.0**	**1.5**	0	12	12	405	39	444	0.05	5

N = 50

	Delta	Alpha	Low	High	Total	Low	High	Total	Min 50	Max 50
Item 1	**-1.8**	**0.5**	0	46	46	3	213	216	0.56	3.62
Item 2	**-1.0**	**0.5**	0	20	20	0	179	179	0.66	4.55
Item 3	**-0.4**	**0.5**	0	17	17	0	147	147	0.66	3.27
Item 4	**0.4**	**0.5**	0	21	21	0	143	143	0.63	2.43
Item 5	**1.0**	**0.5**	0	22	22	1	166	167	0.55	2.74
Item 6	**1.6**	**0.5**	0	30	30	1	180	181	0.55	2.79
Item 7	**-1.6**	**1**	0	10	10	101	47	148	0.27	4.41
Item 8	**-1.2**	**1**	0	13	13	38	39	77	0.41	2.96
Item 9	**-0.6**	**1**	0	3	3	8	26	34	0.41	1.96
Item 10	**-0.2**	**1**	0	2	2	5	20	25	0.49	2.19
Item 11	**0.2**	**1**	0	1	1	3	20	23	0.55	2.1
Item 12	**0.6**	**1**	0	7	7	8	33	41	0.54	3.22
Item 13	**1.2**	**1**	0	12	12	39	36	75	0.38	2.48
Item 14	**1.8**	**1**	0	16	16	129	65	194	0.35	3.45
Item 15	**-1.4**	**1.5**	0	2	2	324	11	335	0.24	2.33
Item 16	**-0.8**	**1.5**	0	1	1	110	7	117	0.38	1.87

Item 17	**0.0**	**1.5**	0	1	1	28	3	31	0.48	2.29
Item 18	**0.8**	**1.5**	0	1	1	94	7	101	0.38	2.13
Item 19	**1.4**	**1.5**	0	2	2	305	11	316	0.27	2.61
Item 20	**2.0**	**1.5**	15	8	23	518	28	546	0.04	2.61

N = 70

	Delta	**Alpha**	**Low**	**High**	**Total**	**Low**	**High**	**Total**	**Min 70**	**Max 70**
Item 1	**-1.8**	**0.5**	0	45	45	1	240	241	0.63	3.16
Item 2	**-1.0**	**0.5**	0	22	22	0	186	186	0.67	2.94
Item 3	**-0.4**	**0.5**	0	16	16	0	208	208	0.74	2.33
Item 4	**0.4**	**0.5**	0	22	22	0	186	186	0.70	2.39
Item 5	**1.0**	**0.5**	0	20	20	0	205	205	0.68	2.62
Item 6	**1.6**	**0.5**	0	41	41	2	228	230	0.58	3.71
Item 7	**-1.6**	**1**	1	12	13	126	57	183	0.26	3.05
Item 8	**-1.2**	**1**	0	6	6	56	36	92	0.51	3.12
Item 9	**-0.6**	**1**	0	2	2	7	21	28	0.60	2.36
Item 10	**-0.2**	**1**	0	4	4	12	26	38	0.58	1.93
Item 11	**0.2**	**1**	0	2	2	8	21	29	0.63	1.89
Item 12	**0.6**	**1**	0	3	3	12	23	35	0.59	2.43
Item 13	**1.2**	**1**	0	2	2	58	33	91	0.47	1.94
Item 14	**1.8**	**1**	1	10	11	143	53	196	0.28	4.39
Item 15	**-1.4**	**1.5**	0	1	1	433	10	443	0.29	1.93
Item 16	**-0.8**	**1.5**	0	1	1	193	4	197	0.49	2.46
Item 17	**0.0**	**1.5**	0	1	1	82	5	87	0.55	1.74
Item 18	**0.8**	**1.5**	0	0	0	207	2	209	0.45	1.64
Item 19	**1.4**	**1.5**	3	4	7	432	18	450	0.21	3.00
Item 20	**2.0**	**1.5**	32	7	39	619	24	643	0.12	2.92

Table A6. Type I Error Rates for Infit Mean Squares (per 1,000 Samples), and Minimum and Maximum Values

Data simulated with multiple discrimination values (Alpha)

				N = 30		*Infit Values*			*N = 50*		*Infit Values*			*N = 70*		*Infit Values*	
	Delta	**Alpha**	**Low**	**High**	**Total**	**Min30**	**Max30**	**Low**	**High**	**Total**	**Min50**	**Max50**	**Low**	**High**	**Total**	**Min70**	**Max**
Item 1	**-1.8**	**0.5**	1	87	88	0.6	1.79	1	99	100	0.71	1.57	0	126	126	0.76	
Item 2	**-1.0**	**0.5**	0	76	76	0.67	1.72	0	108	108	0.76	1.58	0	119	119	0.76	
Item 3	**-0.4**	**0.5**	1	82	83	0.63	1.74	0	111	111	0.74	1.62	0	148	148	0.81	1.
Item 4	**0.4**	**0.5**	1	103	104	0.56	1.99	0	102	102	0.73	1.74	1	135	136	0.73	1.
Item 5	**1.0**	**0.5**	2	99	101	0.59	1.84	0	106	106	0.74	1.63	0	133	133	0.81	1.
Item 6	**1.6**	**0.5**	0	89	89	0.65	1.86	1	118	119	0.67	1.64	0	138	138	0.77	1.
Item 7	**-1.6**	**1**	30	7	37	0.51	1.57	37	3	40	0.51	1.37	45	3	48	0.63	1.
Item 8	**-1.2**	**1**	13	10	23	0.56	1.68	24	5	29	0.61	1.38	26	4	30	0.67	1.
Item 9	**-0.6**	**1**	4	9	13	0.55	1.48	12	5	17	0.51	1.39	10	5	15	0.71	1.
Item 10	**-0.2**	**1**	12	6	18	0.57	1.44	10	5	15	0.59	1.31	18	4	22	0.67	1.
Item 11	**0.2**	**1**	7	7	14	0.58	1.68	11	8	19	0.62	1.36	13	9	22	0.68	1.
Item 12	**0.6**	**1**	11	15	26	0.54	1.55	8	6	14	0.66	1.33	13	3	16	0.65	1.
Item 13	**1.2**	**1**	17	9	26	0.45	1.58	14	3	17	0.57	1.32	21	2	23	0.69	1.
Item 14	**1.8**	**1**	20	12	32	0.48	1.62	25	5	30	0.61	1.37	37	4	41	0.63	
Item 15	**-1.4**	**1.5**	30	7	37	0.51	1.57	37	3	40	0.51	1.37	45	3	48	0.63	1.
Item 16	**-0.8**	**1.5**	13	10	23	0.56	1.68	24	5	29	0.61	1.38	26	4	30	0.67	1.
Item 17	**0.0**	**1.5**	4	9	13	0.55	1.48	12	5	17	0.51	1.39	10	5	15	0.71	1.
Item 18	**0.8**	**1.5**	12	6	18	0.57	1.44	10	5	15	0.59	1.31	18	4	22	0.67	1.
Item 19	**1.4**	**1.5**	7	7	14	0.58	1.68	11	8	19	0.62	1.36	13	9	22	0.68	1.
Item 20	**2.0**	**1.5**	11	15	26	0.54	1.55	8	6	14	0.66	1.33	13	3	16	0.65	1.