Statistical Indexes for Monitoring Item Behavior under Computer Adaptive Testing Environment

Renbang Zhu

Feng Yu

Su Liu

Educational Testing Service

Paper presented at the annual meeting of the American Educational Research Association
New Orleans, April 2002

Abstract

A CAT administration usually requires a large supply of items with accurately estimated psychometric properties, such as IRT parameter estimates, to ensure the precision of examinee ability estimation. However, an estimated IRT model of a given item in any given pool does not always correctly capture what underlies actual examinee responses. This so-called model-data deviation could seriously jeopardize the quality of a test in practice. Therefore, monitoring item behavior in a timely manner is extremely important in CAT for practitioners to take appropriate actions, such as blocking problematic items from active use or pulling the items from subsequent item pools. The purpose of this study was to develop and test two statistical indexes for identifying problematic items with serious model-data deviations. Preliminary results from the simulation study suggested that the new indexes Z_2 and Z_3 could be applied to items with either uniform or non-uniform deviations. Also, results showed that the new indexes exhibited a desired feature. That is, the measured index value monotonically increased as the degree of model-data deviation increased. Further, index Z_3 was more stable across different ability distributions than index Z_2. However, the results indicated that both indexes were sensitive to the variation of examinee sample size.

Introduction

In the context of computer adaptive testing (CAT), an examinee's ability is estimated successively by analyzing the examinee's correct or incorrect response to each of a collection of adaptive items as defined by test specifications. While so many factors can affect the degree of precision in the ability estimation process, the accuracy of the estimation relies heavily on the psychometric property of each delivered item, especially under CAT environment with a relative short test length.

A CAT administration normally requires a large supply of items with accurately estimated psychometric properties in order to sustain its continuous testing. However, a pre-estimated IRT model, which is normally obtained during the process of pretest data analysis, doesn't always correctly capture what underlies a new set of examinee responses to the item. This so called model-data deviation could be caused by many reasons, such as not perfect initial pretest calibration due to estimation methodology or limited calibration sample size, item compromise, differences in motivation of the test takers between the pretest and on-line stage, changes in examinees' learning experience, and so on. The deviation of parameter estimates based on a pre-selected model from what underlying real data could seriously jeopardize the quality of a test.

In the past decade, the concern about the negative impact of model-data deviation has led to the development of statistical procedures and indexes to measure the extent of model-data deviation, measured by the area between a previously estimated item response function and the corresponding newly estimated item response function. The methods include Lagrange Multiplier (LM) test statistic (Glas, 1998, 1999) and the Cumulative Sum (CUSUM) statistic (Glas, 1999; Veerkamp, 1996). The LM test statistic has the advantage of known asymptotic chi-square distribution while the cut value for the CUSUM statistic is related to the desired detection rate in the practical situations (Glas, 1999). Both indexes require item re-calibration, which could be a big challenge for CAT programs. Unlike paper-and-pencil tests, a CAT item is not delivered to all test takers, but rather targeted to examinees within a narrow range of ability levels. As a result, in order to identify misfit items using the above-mentioned methods, it might take a long period of time to accumulate a sufficiently large number of examinees with a wide range of ability, which is quite inconvenient, and sometimes impossible, in the practical settings. Furthermore, large CAT programs usually assemble item pools several months ahead of test administration. Some attractive items (e.g., items with high information) could appear in multiple pre-developed pools. Early detection of items with substantial model-data deviations, especially those compromised items, could help testing programs take appropriate early action, such as blocking the problematic items from active use or removing them from subsequent pools. So, monitoring item behavior in a timely fashion becomes an extremely important practical issue for the CAT programs. The above-mentioned methods do not seem to meet this special need.

In order to identify an item with considerable amount of model-data deviation without going through re-calibration, it is desired to find simple and accurate ways to measure the deviation between the observed examinee response function (ORF) and the corresponding estimated response function (ERF), assuming a pre-selected item response model. For this purpose, Wang et al. developed a statistical index, Z_c, for sequential monitoring of item performance in a CAT operation (Wang, Wingersky, Steffen & Zhu, 1998). The computation of Z_c statistic for a given item i is given by,

$$Z_c = \frac{O_i - E_i}{\sqrt{V_i}} \qquad (1)$$

where

$$O_i = \sum_{j=1}^{N_i} \mu_{ij} \ ,$$

μ_{ij} is the correct response of examinee j to item i, and N_i is the total number of examinees responding to item i;

$$E_i = \sum_{j=1}^{N_i} P_i(\hat{\theta}_j) \ ,$$

$P_i(\hat{\theta}_j)$ is the 3-P logistic function and $\hat{\theta}_j$ is the CAT estimated ability for examinee j; and

$$V_i = \sum_{j=1}^{N_i} P_i(\hat{\theta}_j)[1 - P_i(\hat{\theta}_j)] .$$

As shown above, index Z_c is computed at the item level. It standardizes the overall difference between the observed and the expected total numbers of rights among all examinees responding to the item. According to Smith, Wang, Wingersky & Zhao, (2001), two rules are proposed for computing the Z_c index. The first rule requires that a constant examinee sample size of 400 be used to calculate each Z_c for each item. The rationales for this arrangement are twofold. Number one, using a fixed number of examinees will facilitate the comparison of Z_c statistics from repeated analysis without considering the extraneous effect of sample size variation. Number two, the results from a preliminary study showed that this sample size is reasonably large for a stable misfit estimate but not so large as to limit the number of items that can be statistically evaluated each time. The second rule proposed for computing Z_c is that the examinees included in each analysis sample must fall within the ability range of $[\theta_{\max,i} \pm 1.75]$, where

$$\theta_{\max,i} = b_i + \frac{1}{1.702a_i} \ln \frac{1 + \sqrt{1 + 8c_i}}{2}$$

is the point on the ability scale at which item i yields the maximum information. The reason for setting up this rule is that a CAT item is normally targeted to a specific range of ability. Therefore, it is desirable to minimize measurement error by excluding extreme cases outside this targeted ability range so that the item performance can be evaluated more precisely.

As can be seen in Equation 1, index Z_c is directed to a deviation between the observed overall number of right and the expected overall number of right. It measures an average deviation between the ORF and the ERF curves across a pre-defined ability range. Therefore, index Z_c can be applied to uniform deviation only (Smith, Wang, Wingersky & Zhao, 2001). In other words, this index may work well if an item consistently exhibits easier or harder than expected for examinees at all ability levels. However, empirical data from large-scale CAT programs have shown that a non-uniform deviation often exists. A non-uniform deviation refers to the situation in which the direction of deviation reverses so that an item is differentially harder or easier than expected, conditioned on ability. The Z_c algorithm cannot capture the true model-data deviation under a non-uniform situation, because the effect will cancel each other out if the direction of model deviation changes across different ability levels.

Therefore, it is necessary to expand on Wang, et al. study by looking for a more adequate statistical index to overcome the limitation of Z_c. The proposed index should be simple in computation and easy to implement in practice.

The main purpose of the present study was to evaluate the performances of two new indexes, Z_2 and Z_3, through simulated data. Specifically, the first objective was to investigate whether the new indexes could capture both uniform and non-uniform deviations; and the second objective was to evaluate the sensitivity of the new indexes to such factors as examinee ability distribution and examinee sample size. In the sections that follow, the second section describes the computation of indexes Z_2 and Z_3; the third section discusses simulation design; the fourth section presents the simulation results; and the final section is for conclusions.

Computation of Indexes Z_2 and Z_3

In contrast to the overall difference between the observed and the expected total numbers of right among all examinees within a certain ability range, the computation of Z_2 first classifies examinees into K different ability groups (k = 1, … K), then computes the weighted root-squared difference between the observed and the expected total numbers of right among examinees within the same ability group k, and, finally, sums the weighted differences across all K ability groups. The computation of Z_2 is given as,

$$Z_2 = \frac{\sum_{k=1}^{K} (\frac{n_{ik}}{N_i}) \sqrt{(O_{ik} - E_{ik})^2}}{\sqrt{V_i}}, \qquad (2)$$

where

$$O_{ik} = \sum_{j=1}^{n_{ik}} \mu_{ijk} \ ,$$

μ_{ijk} is the response (0,1) of examinee j in ability group k to item i, n_{ik} is the total number of examinees within ability group k, and N_i is the total number of examinees responding to item i;

and

$$E_{ik} = \sum_{j=1}^{n_{ik}} P_i(\hat{\theta}_{jk}) \,,$$

$P_i(\hat{\theta}_{jk})$ is the 3-P logistic function while $\hat{\theta}_{jk}$ is the CAT estimated ability for examinee j in group k; and V_i is the same as defined in Equation 1.

Index Z_3 employs the conditional error variance within each ability group k, instead of using the grand error variance based on all examinees. In other words, Z_3 first computes the weighted standardized difference within each ability group, then sums across all K ability groups. The computation of Z_3 is shown as follow,

$$Z_3 = \sum_{k=1}^{K} \left(\frac{n_{ik}}{N_i}\right) \sqrt{\frac{(O_{ik} - E_{ik})^2}{V_{ik}}} \,, \tag{3}$$

where

$$V_{ik} = \sum_{j=1}^{n_{ik}} P_{ik}(\hat{\theta}_{jk})(1 - P_{ik}(\hat{\theta}_{jk}) \,,$$

and all other notations in Equation 3 are the same as defined in Equation 2.

Simulation Design

The factors considered in the simulation design were selected to match what commonly occur in practice. The first factor employed was the type of model-data deviation: uniform deviation vs. non-uniform deviation. Figures 1a and 1b present examples of the two types of deviations. As can be seen in the figures, a uniform deviation is merely a shift on the difficulty parameter. Therefore, the observed/real probability of responding an item correctly is consistently larger than the estimated probability, or vice versa. On the other hand, a non-uniform deviation can be caused by a change either on the item discrimination parameter only, or on both difficulty and discrimination parameters. Hence, the estimated probability could be larger than the observed probability at one ability level, but smaller at another level.

The second factor considered in the simulation was examinee ability distribution. Four typical distributions were introduced: normal, rectangular, positively skewed, and negatively skewed. For the normal distribution, the mean of the distribution was set at each item's $\theta_{\max}$ with a variance around 0.7 so that most simulees would fall in the ability range of [θ_{max} ± 1.75]. For the rectangular distribution, the simulated abilities for each item were uniformly distributed on [θ_{max} ± 1.75]. Furthermore, according to previous researches (Pearson & Please, 1975; Fleishman, 1978), a 'typical' non-normal distribution in psychological data was found to have a skew less than 0.8 and a kurtosis between ±0.6. Therefore, coupled with what really happened in the CAT programs being investigated, two skewed distributions in the current study were simulated with the skew index around 0.5 and –0.5 and the kurtosis index around –0.8. Figures 2a through 2d demonstrate the four ability distributions based on four sets of simulated data.

The third factor in the simulation was examinee sample size. Similar to what applied in the computation of Z_c, examinees included in each analysis sample must fall within the ability range of [θ_{max} ± 1.75]. Three levels of sample size were used in the study: 200, 400, and 1,000.

The last factor was the degree of deviation, which was mathematically defined as the total area between the ORF and the ERF, bounded at [θ_{max} ± 1.75]. Five levels of deviation were simulated: no deviation (area equals 0), small deviation (area ranges from 0.09 to 0.15), medium deviation (area ranges from 0.18 to 0.27), large deviation (area ranges from 0.37 to 0.43), and extremely large deviation (area ranges from 0.45 to 0.73).

In order to serve the purposes of the present study, two sets of item parameters were needed for each simulated item: one set of real parameters for creating the ORF curve and another set of estimated parameters for creating the ERF curve.

Three values of item discrimination parameter (0.45, 0.80, 1.14) and three values of item difficulty parameters (-1.69, 0.44, 2.18) were chosen in such a way that they represented the low, middle, and high percentiles of the distributions that might be found in a typical large scale CAT pool. The value of pseudo-guessing parameter was fixed at 0.22, which corresponded to the average value in that particular pool. The three sets of parameter values were used as the estimated parameter values to generated the ERF curve for each of the 9 baseline items (3×3×1). The real item parameters for the ORF curve were created through altering the estimated parameter values to achieve different types and different degrees of deviation. Table 1 lists 72 simulated items with their estimated and real parameter values, the degree and the type of deviation due to parameter changes.

Then, for each simulated item, the behaviors of proposed indexes were examined under every one of the 12 different simulated conditions (4 ability distribution conditions by 3 sample size conditions). Each condition was replicated 100 times.

Simulation Results

The average values over 100 replications under each simulation condition for the three statistical indexes, Z_c, Z_2 and Z_3 , are reported in Tables 2a to 2c for all 72 simulated items, respectively.

The first thing examined here was the performance of the three indexes under different types of deviation (uniform vs. non-uniform deviation). Figures 3a to 3c display the average values of Z_c, Z_2, and Z_3 over 100 replications for two groups of nine items, with each item being responded by 400 simulees under 4 different ability distribution conditions. The first group of items (item 10 to item 18) exhibited uniform deviation, and the second group of items (item 19 to item 27) showed non-uniform deviation. Items in the first group and items in the second group had similar levels of deviation (see Table 1). As can be seen in Figure 3a, the Z_c values for the nine uniform deviation items range from –1 to –2, but the Z_c values for the nine non-uniform deviation items are much closer to zero. This confirms what has been discussed in the introduction. That is, index Z_c can not capture true deviation under the non-uniform deviation scenario, because the positive and negative deviations at different ability levels cancel each other out. Figures 3b and 3c, on the other hand, show that indexes Z_2 and Z_3 were both performing consistently between the two types of deviation. Similar patterns were found with sample sizes 200 and 1,000.

The main goal of the study was to develop a statistical index to reflect the change of item deviation between the ORF and the corresponding ERF. Therefore, the index should possess such a character that the measured index value monotonically increases as the degree of item deviation increases. Furthermore, since the way in evaluating item performance is based on sequentially obtained test data but does not contain a mechanism to control examinee ability distribution in each analysis sample, examining the stability of a selected index across different ability distributions is very important for practitioners. Figures 4a and 4b show the relationships between the measured index value and the simulated degree of deviation, based on sample size 400. One can see from these figures that Z_2 and Z_3 had very similar performances. Both indexes captured the change of item deviation very well. In other words, under each ability distribution condition, both indexes monotonically increased as the degree of deviation increased. However, the figures also exhibited that the measured index values increased at different paces as the degree of deviation increased, especially when the level of deviation was medium or large (i.e., deviation ≥ 0.37). Although both Z_2 and Z_3 performed nearly identical under the normal and the rectangular distribution conditions, the positively skewed distribution gave slightly larger increment than the normal and the rectangular distributions as the simulated degree of deviation increased. On the other hand, the negatively skewed distribution yielded smaller increment than the normal and the rectangular distributions.

As can be noted in Figures 4a and 4b, the values of Z_2 and Z_3 are on different scales. In order to directly compare their stability across the ability distributions, the measured index values under the normal distribution condition were used as the baseline. The ratio between the other three distributions and the baseline were computed. Figures 4c and 4d plot the ratios for Z_2 and Z_3, respectively. Relatively speaking, index Z_3 demonstrated more stability across the ability distributions than index Z_2, especially at the medium level of deviation. Similar results were found under the conditions of sample sizes 200 and 1,000.

Figures 5a and 5b illustrate the trends of Z_2 and Z_3 based on different sample sizes, respectively, as the degree of deviation increases. The values of Z_2 and Z_3 in the two figures were the averages across four ability distributions and over 100 replications under each simulation condition. As can be seen in the figures, under each sample size condition Z_2 and Z_3 were monotonically increasing as the degree of deviation was getting larger. However, the sample size variable had clear impact on both Z_2 and Z_3. That is, at the same level of deviation, the magnitude of measured Z_2 or Z_3 value increased as the sample size increased. Also, the increment was larger for a larger sample size. This was probably due to the fact that, as the sample size is getting larger, the sum of the conditional total score differences increased much faster than the total error variance (see Equation 2) or the squared conditional total score differences increased much faster than the corresponding conditional total error variance (see Equation 3).

Conclusion

The primary goal of this study was to develop an efficient procedure to assess the seriousness of model-data deviation of test items. Since the procedure is mainly for monitoring item behavior, accuracy in detection and simplicity in implementation are both important features for any proposed statistical procedure/index. Both indexes Z_2 and Z_3 carry over the simplicity feature of Z_c. However, the advantage of indexes Z_2 and Z_3 was that the two new indexes were capable of identifying items with both uniform and non-uniform deviations, while the application of index Z_c was restricted to uniform deviations. Also, the proposed new indexes exhibited the sensitivity of capturing the changes of item model-data deviation.

Further, the simulation results indicated that examinee ability distribution had a slight impact on the stability of indexes Z_2 and Z_3. The normal and the rectangular distributions yielded nearly identical results. The positively skewed distribution and the negatively skewed distribution yielded slightly larger and slightly smaller values, respectively, than the normal and the rectangular distributions. But a good thing shown here was that indexes Z_2 and Z_3 were fairly robust to the variation of examinee ability distributions at the moderate level of model-data deviation, where the cutoff value is likely to be chosen for the purpose of item flagging. Relatively speaking, Z_3 displayed a more stable performance across different ability distributions than Z_2. In other words, the measured Z_3 value was affected less by the examinee ability distribution if other conditions remained the same.

No doubt, a substantial impact of examinee sample size on the performance of Z_2 and Z_3 was identified. That is, the measured index value increased as the sample size increased while other conditions were the same. This may be well due to the fact that the increments of the numerator and the denominator in either Equation 2 or Equation 3 are not on the same pace as the examinee sample size increase. Consequently, the results across different sample sizes are not comparable. Clearly, future study should be devoted more to the sensitivity of the indexes to the sample size.

The cutoff point of a selected index is a matter of professional judgment. On one hand, using a too large cut value would run the risk of letting items with serious model- data deviation undetected. On the other hand, if the selected cut point is too small, many items with small or moderate amounts of model-data deviation would be over flagged. Thus, any CAT program should weight carefully between the power factor (i.e., flagging items which should be flagged) and the labor factor (i.e., using more human review time due to over flagging items which should not be flagged) in choosing its cut value.

References

Glas, C.A.W. (1998) Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8, vol. 1*. 647-667.

Glas, C.A.W. (1999) Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64,* 273-294.

Glas, C.A.W. (1999) Item calibration and parameter drift. In Van der Linden, W.J. & Glas, C.A.W. (Eds.): Computerized adaptive testing: theory and practice, 183-199. Netherlands: Kluwer.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. Psychometrika, *43,* 521-532.

Pearson, E.S., & Please, N.W. (1975). Relation between the shape of population distribution of four simple test statistics. *Biometrika, 62*, 223-241.

Smith, R.L., Wang, M.M., Wingersky, M., & Zhao, C. (2001). Monitoring items for changes in performance in computerized adaptive tests. Paper presented at the annual conference of the National Council on Measurement in Education, Seattle, Washington.

Wang, M.M., Wingersky, Steffen, M. & Zhu, R. Preliminary operational item monitoring procedures. Unpublished manuscript.

Verkamp, W.J.J. (1996). Statistical methods for computerized adaptive testing, published doctoral thesis, Twente University, the Netherlands.

Table 1 72 Simulated Items, with Different Types and Various Degrees of Model-data Deviation

Item #	Est_a	Est_b	Est_c	Real_a	Real_b	Real_c	Diff_a	Diff_b	Misfit Area	Misfit Type
1	0.45	-1.69	0.22	0.45	-1.69	0.22	0.00	0.00	0.00	
2	0.45	0.44	0.22	0.45	0.44	0.22	0.00	0.00	0.00	
3	0.45	2.18	0.22	0.45	2.18	0.22	0.00	0.00	0.00	
4	0.80	-1.69	0.22	0.80	-1.69	0.22	0.00	0.00	0.00	
5	0.80	0.44	0.22	0.80	0.44	0.22	0.00	0.00	0.00	
6	0.80	2.18	0.22	0.80	2.18	0.22	0.00	0.00	0.00	
7	1.14	-1.69	0.22	1.14	-1.69	0.22	0.00	0.00	0.00	
8	1.14	0.44	0.22	1.14	0.44	0.22	0.00	0.00	0.00	
9	1.14	2.18	0.22	1.14	2.18	0.22	0.00	0.00	0.00	
10	0.45	-1.69	0.22	0.45	-1.49	0.22	0.00	0.20	0.09	Uniform
11	0.45	0.44	0.22	0.45	0.64	0.22	0.00	0.20	0.09	Uniform
12	0.45	2.18	0.22	0.45	2.38	0.22	0.00	0.20	0.09	Uniform
13	0.80	-1.69	0.22	0.80	-1.49	0.22	0.00	0.20	0.13	Uniform
14	0.80	0.44	0.22	0.80	0.64	0.22	0.00	0.20	0.13	Uniform
15	0.80	2.18	0.22	0.80	2.38	0.22	0.00	0.20	0.13	Uniform
16	1.14	-1.69	0.22	1.14	-1.49	0.22	0.00	0.20	0.15	Uniform
17	1.14	0.44	0.22	1.14	0.64	0.22	0.00	0.20	0.15	Uniform
18	1.14	2.18	0.22	1.14	2.38	0.22	0.00	0.20	0.15	Uniform
19	0.45	-1.69	0.22	0.35	-1.69	0.22	-0.10	0.00	0.09	Non-Uniform
20	0.45	0.44	0.22	0.35	0.44	0.22	-0.10	0.00	0.09	Non-Uniform
21	0.45	2.18	0.22	0.35	2.18	0.22	-0.10	0.00	0.09	Non-Uniform
22	0.80	-1.69	0.22	0.60	-1.69	0.22	-0.20	0.00	0.13	Non-Uniform
23	0.80	0.44	0.22	0.60	0.44	0.22	-0.20	0.00	0.13	Non-Uniform
24	0.80	2.18	0.22	0.60	2.18	0.22	-0.20	0.00	0.13	Non-Uniform
25	1.14	-1.69	0.22	0.84	-1.69	0.22	-0.30	0.00	0.14	Non-Uniform
26	1.14	0.44	0.22	0.84	0.44	0.22	-0.30	0.00	0.14	Non-Uniform
27	1.14	2.18	0.22	0.84	2.18	0.22	-0.30	0.00	0.14	Non-Uniform
28	0.45	-1.29	0.22	0.45	-1.69	0.22	0.00	-0.40	0.18	Uniform
29	0.45	0.84	0.22	0.45	0.44	0.22	0.00	-0.40	0.18	Uniform
30	0.45	2.58	0.22	0.45	2.18	0.22	0.00	-0.40	0.18	Uniform
31	0.80	-1.29	0.22	0.80	-1.69	0.22	0.00	-0.40	0.25	Uniform
32	0.80	0.84	0.22	0.80	0.44	0.22	0.00	-0.40	0.25	Uniform
33	0.80	2.58	0.22	0.80	2.18	0.22	0.00	-0.40	0.25	Uniform
34	0.94	-1.29	0.22	0.94	-1.69	0.22	0.00	-0.40	0.27	Uniform
35	0.94	0.84	0.22	0.94	0.44	0.22	0.00	-0.40	0.27	Uniform
36	0.94	2.58	0.22	0.94	2.18	0.22	0.00	-0.40	0.27	Uniform
37	0.45	-1.69	0.22	0.60	-1.34	0.22	0.15	0.35	0.17	Non-Uniform
38	0.45	0.44	0.22	0.60	0.79	0.22	0.15	0.35	0.17	Non-Uniform
39	0.45	2.18	0.22	0.60	2.53	0.22	0.15	0.35	0.17	Non-Uniform
40	0.80	-1.69	0.22	1.11	-1.34	0.22	0.31	0.35	0.25	Non-Uniform
41	0.80	0.44	0.22	1.11	0.79	0.22	0.31	0.35	0.25	Non-Uniform
42	0.80	2.18	0.22	1.11	2.53	0.22	0.31	0.35	0.25	Non-Uniform
43	1.14	-1.69	0.22	1.58	-1.34	0.22	0.44	0.35	0.27	Non-Uniform
44	1.14	0.44	0.22	1.58	0.79	0.22	0.44	0.35	0.27	Non-Uniform
45	1.14	2.18	0.22	1.58	2.53	0.22	0.44	0.35	0.27	Non-Uniform
46	0.45	-1.69	0.22	0.45	-2.29	0.22	0.00	-0.60	0.26	Uniform
47	0.45	0.44	0.22	0.45	-0.16	0.22	0.00	-0.60	0.26	Uniform
48	0.45	2.18	0.22	0.45	1.58	0.22	0.00	-0.60	0.26	Uniform
49	0.80	-1.69	0.22	0.80	-2.29	0.22	0.00	-0.60	0.37	Uniform
50	0.80	0.44	0.22	0.80	-0.16	0.22	0.00	-0.60	0.37	Uniform
51	0.80	2.18	0.22	0.80	1.58	0.22	0.00	-0.60	0.37	Uniform
52	1.14	-1.69	0.22	1.14	-2.29	0.22	0.00	-0.60	0.43	Uniform
53	1.14	0.44	0.22	1.14	-0.16	0.22	0.00	-0.60	0.43	Uniform
54	1.14	2.18	0.22	1.14	1.58	0.22	0.00	-0.60	0.43	Uniform
55	0.80	-1.69	0.22	0.50	-2.19	0.22	-0.30	-0.50	0.26	Non-Uniform
56	0.80	0.44	0.22	0.50	-0.06	0.22	-0.30	-0.50	0.26	Non-Uniform
57	0.80	2.18	0.22	0.50	1.68	0.22	-0.30	-0.50	0.26	Non-Uniform
58	0.80	-1.69	0.22	0.35	-2.49	0.22	-0.45	-0.80	0.37	Non-Uniform
59	0.80	0.44	0.22	0.35	-0.36	0.22	-0.45	-0.80	0.37	Non-Uniform
60	0.80	2.18	0.22	0.35	1.38	0.22	-0.45	-0.80	0.37	Non-Uniform
61	0.80	-1.69	0.22	0.30	-2.79	0.22	-0.50	-1.10	0.43	Non-Uniform
62	0.80	0.44	0.22	0.30	-0.66	0.22	-0.50	-1.10	0.43	Non-Uniform
63	0.80	2.18	0.22	0.30	1.08	0.22	-0.50	-1.10	0.43	Non-Uniform
64	0.80	-1.69	0.22	0.40	-2.89	0.22	-0.40	-1.20	0.45	Non-Uniform
65	0.80	0.44	0.22	0.40	-0.76	0.22	-0.40	-1.20	0.45	Non-Uniform
66	0.80	2.18	0.22	0.40	0.98	0.22	-0.40	-1.20	0.45	Non-Uniform
67	0.80	-1.69	0.22	0.40	-3.09	0.22	-0.40	-1.40	0.51	Non-Uniform
68	0.80	0.44	0.22	0.40	-0.96	0.22	-0.40	-1.40	0.51	Non-Uniform
69	0.80	2.18	0.22	0.40	0.78	0.22	-0.40	-1.40	0.51	Non-Uniform
70	0.80	-1.69	0.22	0.40	-3.50	0.22	-0.40	-1.81	0.62	Non-Uniform
71	0.80	0.44	0.22	0.40	-1.50	0.22	-0.40	-1.94	0.66	Non-Uniform
72	0.80	2.18	0.22	0.40	0.00	0.22	-0.40	-2.18	0.73	Non-Uniform

Table 2a **Mean for Z_c, Z_2, and Z_3 over 100 Replications under 4 Ability Distribution Conditions (Sample Size = 200)**

Item #	Area	Z_c_rct	Z_c_nom	Z_c _pos	Z_c _neg	Z_2_rct	Z_2_nom	Z_2_pos	Z_2_neg	Z_3_rct	Z_3_nom	Z_3_pos	Z_3_neg
1	**0.00**	0.04	-0.11	0.04	0.09	0.30	0.33	0.33	0.30	0.80	0.83	0.84	0.78
2	**0.00**	0.16	0.06	0.22	0.04	0.30	0.32	0.32	0.33	0.79	0.80	0.81	0.86
3	**0.00**	-0.10	0.08	0.05	0.09	0.29	0.31	0.32	0.30	0.77	0.77	0.81	0.79
4	**0.00**	0.15	-0.19	0.09	-0.08	0.30	0.34	0.32	0.30	0.79	0.82	0.80	0.79
5	**0.00**	0.00	0.05	-0.11	0.03	0.30	0.31	0.33	0.30	0.79	0.79	0.82	0.80
6	**0.00**	0.07	-0.01	0.01	-0.01	0.30	0.32	0.30	0.31	0.81	0.80	0.76	0.84
7	**0.00**	0.09	0.06	0.09	0.11	0.31	0.32	0.30	0.29	0.83	0.80	0.76	0.81
8	**0.00**	-0.16	-0.08	-0.04	0.02	0.28	0.32	0.31	0.28	0.77	0.80	0.79	0.78
9	**0.00**	0.10	0.07	-0.01	0.06	0.30	0.32	0.32	0.28	0.81	0.81	0.81	0.79
10	**0.09**	-0.85	-0.62	-0.82	-0.86	0.33	0.33	0.33	0.33	0.87	0.83	0.82	0.85
11	**0.09**	-0.85	-0.92	-0.82	-0.86	0.31	0.35	0.33	0.32	0.82	0.87	0.83	0.84
12	**0.09**	-0.68	-0.80	-0.70	-0.83	0.31	0.35	0.33	0.34	0.82	0.88	0.84	0.87
13	**0.13**	-1.23	-1.38	-1.27	-1.40	0.35	0.40	0.36	0.36	0.94	0.97	0.89	0.96
14	**0.13**	-1.18	-1.36	-1.15	-1.27	0.32	0.39	0.36	0.37	0.86	0.96	0.91	0.98
15	**0.13**	-1.16	-1.36	-1.15	-1.31	0.33	0.39	0.35	0.36	0.90	0.97	0.89	0.97
16	**0.15**	-1.57	-1.84	-1.36	-1.75	0.35	0.44	0.37	0.38	0.96	1.08	0.95	1.04
17	**0.15**	-1.53	-1.68	-1.42	-1.53	0.36	0.43	0.37	0.36	0.98	1.05	0.94	1.00
18	**0.15**	-1.30	-1.81	-1.43	-1.33	0.36	0.44	0.37	0.35	0.98	1.08	0.96	0.98
19	**0.09**	-0.34	-0.36	0.02	-0.54	0.33	0.33	0.33	0.34	0.88	0.82	0.83	0.87
20	**0.09**	-0.34	-0.21	0.03	-0.50	0.32	0.34	0.32	0.36	0.85	0.85	0.82	0.94
21	**0.09**	-0.22	-0.25	0.08	-0.60	0.33	0.32	0.35	0.35	0.89	0.81	0.88	0.90
22	**0.13**	-0.24	-0.43	0.22	-0.65	0.36	0.35	0.37	0.36	0.98	0.90	0.94	0.99
23	**0.13**	-0.31	-0.41	0.12	-0.47	0.36	0.34	0.36	0.37	0.98	0.87	0.94	1.00
24	**0.13**	0.09	-0.13	0.11	-0.74	0.34	0.35	0.36	0.37	0.91	0.90	0.93	1.02
25	**0.14**	0.02	0.02	0.37	-0.39	0.37	0.36	0.38	0.39	1.08	0.92	0.99	1.16
26	**0.14**	-0.22	-0.10	0.12	-0.68	0.39	0.40	0.37	0.37	1.12	1.04	0.98	1.12
27	**0.14**	0.04	-0.29	0.16	-0.66	0.34	0.38	0.36	0.39	0.96	0.99	0.96	1.17
28	**0.18**	1.65	1.49	1.51	1.55	0.34	0.37	0.38	0.35	0.90	0.91	0.93	0.91
29	**0.18**	1.55	1.44	1.39	1.61	0.35	0.39	0.37	0.35	0.90	0.95	0.92	0.90
30	**0.18**	1.48	1.62	1.63	1.58	0.34	0.39	0.39	0.36	0.91	0.96	0.96	0.92
31	**0.25**	2.32	2.51	2.49	2.35	0.41	0.50	0.48	0.40	1.07	1.17	1.16	1.05
32	**0.25**	2.48	2.36	2.34	2.24	0.43	0.47	0.44	0.41	1.13	1.14	1.08	1.06
33	**0.25**	2.38	2.65	2.44	2.37	0.41	0.52	0.46	0.40	1.07	1.23	1.11	1.06
34	**0.27**	2.61	2.88	2.59	2.50	0.45	0.54	0.49	0.41	1.17	1.29	1.17	1.07
35	**0.27**	2.63	2.73	2.72	2.73	0.45	0.53	0.50	0.45	1.16	1.25	1.20	1.18
36	**0.27**	2.59	2.92	2.79	2.63	0.43	0.56	0.50	0.43	1.12	1.30	1.21	1.12
37	**0.17**	-1.36	-1.49	-1.75	-0.99	0.38	0.40	0.44	0.35	0.98	0.97	1.05	0.90
38	**0.17**	-1.44	-1.38	-1.72	-1.05	0.37	0.38	0.42	0.33	0.95	0.92	1.03	0.85
39	**0.17**	-1.52	-1.31	-1.66	-1.04	0.40	0.37	0.42	0.35	1.02	0.91	1.02	0.90
40	**0.25**	-2.30	-2.33	-2.43	-1.91	0.44	0.52	0.50	0.40	1.12	1.21	1.21	1.02
41	**0.25**	-2.39	-2.37	-2.66	-1.88	0.46	0.52	0.52	0.39	1.16	1.21	1.24	1.02
42	**0.25**	-2.29	-2.41	-2.73	-1.92	0.45	0.54	0.54	0.40	1.16	1.25	1.29	1.03
43	**0.27**	-2.64	-3.13	-2.84	-2.37	0.48	0.64	0.55	0.45	1.21	1.45	1.33	1.17
44	**0.27**	-2.68	-3.26	-2.92	-2.35	0.48	0.68	0.55	0.45	1.22	1.53	1.34	1.15
45	**0.27**	-2.63	-3.13	-2.72	-2.43	0.48	0.66	0.52	0.43	1.22	1.50	1.25	1.09
46	**0.26**	2.40	2.29	2.39	2.33	0.41	0.46	0.46	0.41	1.07	1.10	1.13	1.06
47	**0.26**	2.48	2.25	2.33	2.35	0.42	0.45	0.45	0.42	1.09	1.10	1.10	1.09
48	**0.26**	2.28	2.29	2.19	2.28	0.41	0.46	0.46	0.41	1.06	1.10	1.11	1.06
49	**0.37**	3.41	3.68	3.85	3.48	0.53	0.67	0.66	0.53	1.37	1.57	1.57	1.39
50	**0.37**	3.35	3.71	3.71	3.43	0.53	0.66	0.64	0.51	1.36	1.55	1.52	1.33
51	**0.37**	3.49	3.67	3.72	3.39	0.54	0.66	0.65	0.51	1.38	1.54	1.54	1.32
52	**0.43**	4.28	4.72	4.41	4.17	0.65	0.88	0.76	0.60	1.62	1.96	1.77	1.55
53	**0.43**	4.27	4.69	4.34	4.21	0.64	0.85	0.74	0.59	1.60	1.93	1.73	1.53
54	**0.43**	4.10	4.62	4.36	4.01	0.61	0.84	0.75	0.57	1.54	1.89	1.75	1.48
55	**0.26**	2.15	1.77	2.56	1.22	0.48	0.46	0.61	0.39	1.23	1.12	1.42	1.03
56	**0.26**	1.96	2.03	2.78	1.31	0.46	0.48	0.64	0.40	1.18	1.16	1.50	1.09
57	**0.26**	1.98	2.13	2.69	1.35	0.47	0.50	0.62	0.40	1.22	1.21	1.43	1.07
58	**0.37**	2.29	2.16	3.14	0.98	0.60	0.55	0.78	0.53	1.57	1.36	1.80	1.46
59	**0.37**	2.16	2.08	3.29	1.20	0.58	0.53	0.78	0.50	1.49	1.30	1.81	1.38
60	**0.37**	2.14	1.95	3.19	0.92	0.59	0.53	0.79	0.50	1.53	1.32	1.83	1.38
61	**0.43**	2.58	2.74	3.70	1.49	0.65	0.61	0.89	0.54	1.71	1.51	2.03	1.49
62	**0.43**	2.73	2.39	3.78	1.51	0.65	0.59	0.89	0.56	1.70	1.47	2.03	1.54
63	**0.43**	2.81	2.38	3.86	1.46	0.66	0.58	0.89	0.55	1.69	1.45	2.04	1.51
64	**0.45**	4.10	3.89	4.95	2.96	0.69	0.72	0.97	0.51	1.70	1.69	2.19	1.33
65	**0.45**	3.92	3.75	4.91	3.18	0.68	0.70	0.97	0.52	1.68	1.64	2.18	1.35
66	**0.45**	3.92	3.71	4.88	2.97	0.69	0.68	0.97	0.51	1.72	1.60	2.17	1.34
67	**0.51**	4.62	4.54	5.33	3.92	0.75	0.80	1.04	0.56	1.85	1.85	2.34	1.45
68	**0.51**	4.82	4.53	5.48	3.63	0.76	0.80	1.06	0.55	1.86	1.85	2.35	1.41
69	**0.51**	4.54	4.54	5.53	3.62	0.74	0.78	1.08	0.55	1.83	1.83	2.40	1.42
70	**0.62**	5.93	5.60	6.85	4.87	0.90	0.96	1.28	0.66	2.19	2.21	2.84	1.69
71	**0.66**	6.30	6.07	7.04	4.93	0.95	1.03	1.30	0.67	2.31	2.38	2.90	1.71
72	**0.73**	6.84	6.56	7.83	5.82	1.02	1.10	1.43	0.77	2.49	2.55	3.17	1.97

Table 2b Mean for Z_c, Z_2, and Z_3 over 100 Replications under 4 Ability Distribution Conditions (Sample Size = 400)

Item #	Area	Z_c_rct	Z_c_nom	Z_c _pos	Z_c _neg	Z_2_rct	Z_2_nom	Z_2_pos	Z_2_neg	Z_3_rct	Z_3_nom	Z_3_pos	Z_3_neg
1	**0.00**	-0.01	0.16	-0.04	0.02	0.31	0.31	0.30	0.33	0.81	0.78	0.77	0.86
2	**0.00**	-0.07	-0.06	0.05	-0.05	0.29	0.33	0.30	0.29	0.78	0.81	0.76	0.75
3	**0.00**	0.03	0.03	-0.09	-0.06	0.31	0.33	0.31	0.29	0.81	0.82	0.78	0.75
4	**0.00**	0.08	0.10	0.13	-0.01	0.30	0.33	0.31	0.30	0.81	0.83	0.78	0.81
5	**0.00**	0.13	-0.06	-0.16	-0.08	0.28	0.33	0.31	0.30	0.76	0.82	0.78	0.80
6	**0.00**	-0.05	0.04	0.10	0.16	0.29	0.31	0.31	0.30	0.78	0.77	0.78	0.80
7	**0.00**	-0.03	0.04	0.05	0.13	0.29	0.34	0.30	0.27	0.77	0.85	0.77	0.77
8	**0.00**	-0.07	0.09	-0.08	-0.11	0.28	0.31	0.32	0.28	0.78	0.78	0.82	0.78
9	**0.00**	-0.21	0.08	0.10	-0.11	0.29	0.30	0.31	0.28	0.81	0.76	0.78	0.80
10	**0.09**	-1.00	-1.28	-1.38	-1.08	0.34	0.37	0.37	0.33	0.90	0.92	0.92	0.86
11	**0.09**	-1.09	-1.08	-1.09	-1.18	0.32	0.35	0.35	0.36	0.84	0.88	0.88	0.92
12	**0.09**	-1.04	-1.09	-1.30	-1.02	0.34	0.37	0.35	0.34	0.89	0.91	0.89	0.89
13	**0.13**	-1.71	-1.88	-1.60	-1.76	0.37	0.43	0.37	0.39	1.00	1.04	0.95	1.05
14	**0.13**	-1.83	-1.81	-1.83	-1.99	0.38	0.43	0.39	0.40	1.03	1.04	0.99	1.07
15	**0.13**	-1.70	-1.93	-1.76	-1.90	0.36	0.44	0.39	0.41	0.98	1.07	0.99	1.09
16	**0.15**	-2.01	-2.40	-2.00	-2.03	0.40	0.51	0.42	0.40	1.09	1.24	1.07	1.09
17	**0.15**	-2.08	-2.29	-2.01	-2.15	0.40	0.52	0.41	0.40	1.11	1.25	1.03	1.08
18	**0.15**	-2.05	-2.34	-1.83	-2.24	0.40	0.50	0.39	0.43	1.10	1.22	1.01	1.20
19	**0.09**	-0.38	-0.55	-0.11	-0.66	0.34	0.35	0.36	0.37	0.91	0.89	0.90	0.97
20	**0.09**	-0.40	-0.48	-0.12	-0.85	0.32	0.33	0.32	0.37	0.86	0.84	0.82	0.98
21	**0.09**	-0.25	-0.50	-0.09	-0.74	0.35	0.35	0.35	0.37	0.92	0.87	0.90	0.96
22	**0.13**	-0.32	-0.30	0.25	-0.70	0.36	0.38	0.37	0.42	1.02	0.99	0.95	1.17
23	**0.13**	-0.25	-0.18	0.37	-0.88	0.36	0.39	0.42	0.42	1.03	1.00	1.07	1.17
24	**0.13**	-0.45	-0.42	0.14	-0.97	0.37	0.39	0.37	0.42	1.05	1.02	0.95	1.17
25	**0.14**	-0.07	-0.36	0.39	-0.67	0.38	0.42	0.41	0.46	1.12	1.11	1.10	1.39
26	**0.14**	-0.19	-0.24	0.31	-0.97	0.41	0.42	0.41	0.44	1.21	1.12	1.08	1.36
27	**0.14**	-0.14	-0.39	0.40	-0.82	0.41	0.40	0.43	0.45	1.22	1.09	1.11	1.39
28	**0.18**	2.26	2.20	2.26	2.19	0.41	0.44	0.45	0.41	1.08	1.08	1.11	1.08
29	**0.18**	2.24	2.33	2.23	2.17	0.40	0.46	0.44	0.40	1.04	1.11	1.08	1.03
30	**0.18**	2.35	2.26	2.15	2.19	0.40	0.47	0.42	0.41	1.07	1.14	1.05	1.06
31	**0.25**	3.41	3.75	3.47	3.22	0.52	0.67	0.60	0.51	1.36	1.57	1.46	1.34
32	**0.25**	3.31	3.46	3.42	3.35	0.50	0.63	0.58	0.51	1.31	1.49	1.41	1.35
33	**0.25**	3.43	3.63	3.62	3.35	0.52	0.67	0.62	0.52	1.35	1.56	1.49	1.37
34	**0.27**	3.58	4.11	3.82	3.51	0.55	0.75	0.65	0.53	1.41	1.74	1.57	1.39
35	**0.27**	3.77	3.98	3.79	3.57	0.57	0.72	0.64	0.53	1.49	1.67	1.55	1.39
36	**0.27**	3.73	4.18	3.86	3.44	0.56	0.75	0.65	0.52	1.46	1.75	1.56	1.37
37	**0.17**	-2.16	-2.04	-2.35	-1.45	0.44	0.46	0.52	0.38	1.13	1.12	1.25	0.99
38	**0.17**	-1.90	-1.90	-2.39	-1.44	0.43	0.46	0.53	0.37	1.12	1.11	1.26	0.96
39	**0.17**	-1.97	-2.21	-2.34	-1.52	0.41	0.47	0.51	0.38	1.07	1.13	1.23	1.00
40	**0.25**	-3.11	-3.67	-3.74	-2.63	0.55	0.72	0.69	0.48	1.39	1.64	1.62	1.23
41	**0.25**	-3.14	-3.59	-3.47	-2.49	0.54	0.71	0.67	0.44	1.35	1.65	1.57	1.15
42	**0.25**	-3.13	-3.68	-3.69	-2.87	0.54	0.70	0.69	0.49	1.36	1.62	1.63	1.26
43	**0.27**	-3.69	-4.45	-3.87	-3.53	0.60	0.88	0.68	0.56	1.50	1.94	1.62	1.41
44	**0.27**	-3.70	-4.25	-3.95	-3.49	0.61	0.83	0.70	0.55	1.51	1.87	1.66	1.38
45	**0.27**	-3.68	-4.46	-4.02	-3.49	0.60	0.87	0.71	0.54	1.50	1.94	1.68	1.38
46	**0.26**	3.24	3.27	3.22	3.26	0.50	0.58	0.56	0.53	1.32	1.41	1.37	1.37
47	**0.26**	3.42	3.42	3.30	3.23	0.52	0.61	0.57	0.52	1.36	1.46	1.39	1.35
48	**0.26**	3.30	3.36	3.31	3.09	0.51	0.59	0.57	0.49	1.35	1.42	1.40	1.27
49	**0.37**	4.93	5.25	4.94	4.72	0.72	0.92	0.83	0.68	1.86	2.15	1.97	1.78
50	**0.37**	4.99	5.31	5.13	4.79	0.73	0.94	0.86	0.69	1.88	2.19	2.04	1.80
51	**0.37**	5.01	5.29	5.13	4.78	0.73	0.92	0.86	0.69	1.88	2.15	2.04	1.80
52	**0.43**	5.91	6.55	6.25	5.72	0.86	1.17	1.05	0.79	2.16	2.64	2.44	2.06
53	**0.43**	6.08	6.56	6.29	5.62	0.89	1.18	1.05	0.78	2.22	2.65	2.45	2.03
54	**0.43**	6.23	6.65	6.05	5.79	0.91	1.20	1.01	0.81	2.27	2.69	2.35	2.09
55	**0.26**	2.84	2.88	3.72	1.78	0.58	0.59	0.79	0.46	1.48	1.44	1.80	1.26
56	**0.26**	2.73	2.61	3.73	1.81	0.58	0.56	0.79	0.47	1.50	1.35	1.81	1.27
57	**0.26**	2.82	2.78	3.57	1.68	0.60	0.59	0.78	0.46	1.52	1.41	1.81	1.23
58	**0.37**	3.05	3.03	4.50	1.62	0.77	0.69	1.04	0.63	2.02	1.70	2.39	1.76
59	**0.37**	3.06	3.07	4.33	1.59	0.77	0.70	1.02	0.63	2.02	1.72	2.36	1.77
60	**0.37**	3.13	2.95	4.44	1.55	0.78	0.69	1.04	0.65	2.04	1.69	2.38	1.82
61	**0.43**	3.94	3.55	5.21	2.07	0.87	0.79	1.21	0.72	2.24	1.97	2.76	2.00
62	**0.43**	3.73	3.46	5.22	2.07	0.85	0.77	1.21	0.72	2.20	1.92	2.76	2.02
63	**0.43**	3.75	3.53	5.14	2.13	0.86	0.77	1.19	0.70	2.25	1.89	2.71	1.96
64	**0.45**	5.64	5.56	6.79	4.36	0.91	0.96	1.31	0.66	2.23	2.22	2.93	1.73
65	**0.45**	5.44	5.44	6.85	4.21	0.90	0.95	1.34	0.65	2.22	2.22	2.99	1.70
66	**0.45**	5.64	5.35	6.66	4.35	0.90	0.94	1.30	0.65	2.20	2.21	2.91	1.69
67	**0.51**	6.40	6.29	7.75	5.22	0.99	1.06	1.47	0.72	2.44	2.48	3.26	1.87
68	**0.51**	6.56	6.42	7.79	5.21	1.02	1.08	1.47	0.71	2.51	2.51	3.27	1.84
69	**0.51**	6.65	6.42	7.82	5.23	1.01	1.10	1.48	0.72	2.48	2.55	3.29	1.87
70	**0.62**	8.28	8.13	9.61	6.88	1.24	1.36	1.76	0.88	3.01	3.13	3.91	2.26
71	**0.66**	8.89	8.46	10.27	7.25	1.31	1.40	1.88	0.93	3.19	3.24	4.17	2.38
72	**0.73**	9.77	9.46	11.11	8.06	1.42	1.57	2.02	1.02	3.47	3.62	4.48	2.63

Table 2c Mean for Z_c, Z_2, and Z_3 over 100 Replications under 4 Ability Distribution Conditions (Sample Size = 1000)

Item #	Area	Z_c_rct	Z_c_nom	Z_c _pos	Z_c _neg	Z_2_rct	Z_2_nom	Z_2_pos	Z_2_neg	Z_3_rct	Z_3_nom	Z_3_pos	Z_3_neg
1	**0.00**	-0.25	0.13	0.00	-0.13	0.31	0.34	0.30	0.29	0.82	0.84	0.76	0.77
2	**0.00**	-0.04	0.07	-0.04	0.01	0.30	0.34	0.32	0.29	0.81	0.84	0.81	0.76
3	**0.00**	-0.03	-0.05	0.11	-0.03	0.29	0.34	0.33	0.31	0.78	0.86	0.83	0.81
4	**0.00**	0.00	-0.04	0.07	0.05	0.29	0.32	0.31	0.30	0.80	0.81	0.80	0.80
5	**0.00**	0.07	-0.03	-0.19	-0.12	0.31	0.32	0.34	0.31	0.84	0.80	0.85	0.83
6	**0.00**	0.01	0.05	0.03	-0.18	0.30	0.32	0.31	0.31	0.81	0.82	0.80	0.82
7	**0.00**	0.05	-0.04	-0.11	-0.16	0.28	0.31	0.29	0.29	0.77	0.79	0.75	0.81
8	**0.00**	-0.10	0.16	-0.02	0.05	0.29	0.29	0.33	0.30	0.81	0.74	0.84	0.84
9	**0.00**	-0.10	0.04	0.08	0.21	0.29	0.32	0.31	0.30	0.80	0.80	0.79	0.83
10	**0.09**	-1.93	-1.91	-1.79	-1.63	0.38	0.43	0.41	0.36	1.01	1.06	1.02	0.93
11	**0.09**	-1.81	-2.00	-1.81	-1.77	0.38	0.44	0.38	0.39	1.01	1.08	0.97	1.02
12	**0.09**	-1.66	-1.94	-1.67	-1.88	0.37	0.43	0.38	0.40	0.98	1.07	0.97	1.04
13	**0.13**	-2.73	-3.00	-2.51	-2.84	0.46	0.57	0.46	0.48	1.25	1.38	1.17	1.30
14	**0.13**	-2.86	-2.87	-2.71	-2.77	0.46	0.56	0.50	0.48	1.25	1.34	1.25	1.28
15	**0.13**	-2.67	-2.91	-2.62	-2.79	0.46	0.56	0.47	0.47	1.23	1.36	1.20	1.27
16	**0.15**	-3.20	-3.71	-3.14	-3.55	0.53	0.71	0.54	0.58	1.44	1.69	1.39	1.62
17	**0.15**	-3.20	-3.55	-3.23	-3.50	0.53	0.69	0.56	0.57	1.45	1.64	1.40	1.56
18	**0.15**	-3.19	-4.00	-3.25	-3.30	0.51	0.75	0.56	0.54	1.39	1.77	1.41	1.49
19	**0.09**	-0.57	-0.67	-0.06	-1.16	0.38	0.37	0.38	0.44	1.02	0.94	0.97	1.15
20	**0.09**	-0.50	-0.60	-0.07	-0.98	0.38	0.38	0.40	0.44	1.04	0.97	1.02	1.16
21	**0.09**	-0.86	-0.77	-0.12	-1.26	0.37	0.39	0.39	0.45	1.02	0.99	1.00	1.17
22	**0.13**	-0.42	-0.65	0.46	-1.19	0.48	0.45	0.48	0.55	1.39	1.19	1.25	1.57
23	**0.13**	-0.18	-0.56	0.39	-1.18	0.49	0.45	0.49	0.53	1.40	1.18	1.26	1.52
24	**0.13**	-0.41	-0.32	0.52	-1.52	0.48	0.45	0.52	0.54	1.37	1.22	1.32	1.53
25	**0.14**	-0.26	-0.36	0.67	-1.08	0.52	0.48	0.52	0.61	1.60	1.32	1.40	1.89
26	**0.14**	-0.22	-0.32	0.76	-1.13	0.50	0.49	0.53	0.57	1.56	1.35	1.41	1.80
27	**0.14**	-0.26	-0.38	0.68	-1.22	0.51	0.52	0.53	0.58	1.56	1.43	1.40	1.84
28	**0.18**	3.66	3.69	3.64	3.53	0.55	0.63	0.60	0.57	1.44	1.54	1.49	1.47
29	**0.18**	3.51	3.51	3.60	3.55	0.54	0.62	0.60	0.56	1.41	1.50	1.48	1.46
30	**0.18**	3.51	3.67	3.60	3.46	0.54	0.65	0.60	0.55	1.41	1.57	1.47	1.43
31	**0.25**	5.33	5.86	5.50	5.17	0.77	1.03	0.90	0.75	2.01	2.40	2.17	1.98
32	**0.25**	5.38	5.94	5.51	5.21	0.78	1.03	0.91	0.76	2.02	2.42	2.18	1.99
33	**0.25**	5.34	5.71	5.38	5.33	0.77	0.99	0.87	0.77	2.02	2.32	2.12	2.03
34	**0.27**	5.81	6.32	5.80	5.78	0.84	1.11	0.95	0.83	2.18	2.57	2.28	2.19
35	**0.27**	5.72	6.49	5.86	5.78	0.83	1.15	0.96	0.83	2.15	2.66	2.30	2.19
36	**0.27**	5.60	6.46	5.93	5.61	0.81	1.14	0.96	0.81	2.12	2.64	2.32	2.13
37	**0.17**	-2.99	-3.17	-3.90	-2.45	0.55	0.61	0.77	0.47	1.40	1.48	1.78	1.25
38	**0.17**	-2.96	-3.28	-3.67	-2.36	0.56	0.62	0.74	0.45	1.44	1.49	1.71	1.18
39	**0.17**	-3.10	-3.25	-3.88	-2.39	0.56	0.63	0.78	0.45	1.43	1.52	1.81	1.19
40	**0.25**	-5.17	-5.52	-5.64	-4.25	0.83	1.03	1.02	0.65	2.07	2.33	2.36	1.68
41	**0.25**	-5.00	-5.34	-5.75	-4.28	0.81	1.00	1.04	0.66	2.01	2.26	2.41	1.69
42	**0.25**	-5.05	-5.50	-5.75	-4.36	0.80	1.02	1.03	0.67	2.00	2.31	2.38	1.73
43	**0.27**	-5.79	-7.27	-6.14	-5.41	0.89	1.38	1.04	0.80	2.17	3.00	2.43	1.97
44	**0.27**	-5.85	-6.92	-6.13	-5.32	0.89	1.34	1.04	0.79	2.19	2.91	2.42	1.96
45	**0.27**	-5.88	-7.04	-6.22	-5.50	0.90	1.34	1.05	0.81	2.19	2.92	2.45	1.98
46	**0.26**	5.26	5.34	5.24	5.14	0.76	0.90	0.85	0.77	1.99	2.17	2.08	2.01
47	**0.26**	5.25	5.31	5.30	5.42	0.76	0.90	0.86	0.81	1.99	2.17	2.10	2.10
48	**0.26**	5.07	5.23	5.35	5.10	0.73	0.87	0.87	0.76	1.94	2.11	2.13	1.97
49	**0.37**	7.86	8.25	8.05	7.48	1.13	1.42	1.32	1.06	2.92	3.31	3.15	2.79
50	**0.37**	7.94	8.36	8.07	7.40	1.14	1.45	1.32	1.06	2.95	3.37	3.16	2.78
51	**0.37**	7.86	8.32	8.12	7.55	1.13	1.44	1.34	1.07	2.92	3.35	3.19	2.80
52	**0.43**	9.57	10.53	9.95	9.18	1.38	1.86	1.64	1.25	3.45	4.19	3.83	3.24
53	**0.43**	9.59	10.53	9.94	9.08	1.38	1.88	1.64	1.25	3.46	4.22	3.83	3.23
54	**0.43**	9.35	10.62	9.89	9.30	1.34	1.89	1.63	1.27	3.38	4.25	3.81	3.30
55	**0.26**	4.36	4.30	5.81	2.84	0.84	0.82	1.19	0.62	2.16	1.98	2.71	1.69
56	**0.26**	4.36	4.33	5.85	2.92	0.84	0.82	1.21	0.64	2.13	1.97	2.74	1.76
57	**0.26**	4.31	4.35	5.63	3.01	0.84	0.84	1.18	0.65	2.14	1.99	2.66	1.78
58	**0.37**	4.90	4.66	7.02	2.53	1.15	1.02	1.61	0.95	3.02	2.53	3.66	2.69
59	**0.37**	4.70	4.61	7.08	2.47	1.16	1.03	1.62	0.97	3.07	2.57	3.71	2.73
60	**0.37**	5.01	4.59	7.22	2.59	1.15	1.02	1.61	0.95	3.02	2.54	3.66	2.68
61	**0.43**	5.84	5.68	8.34	2.95	1.31	1.18	1.86	1.06	3.42	2.89	4.21	3.01
62	**0.43**	5.97	5.44	8.56	3.24	1.30	1.15	1.88	1.07	3.37	2.85	4.26	3.00
63	**0.43**	5.95	5.69	8.33	3.13	1.32	1.17	1.84	1.06	3.45	2.87	4.18	3.01
64	**0.45**	9.02	8.65	10.89	6.76	1.40	1.47	2.06	0.99	3.43	3.39	4.57	2.60
65	**0.45**	9.01	8.55	10.89	6.74	1.41	1.45	2.08	0.95	3.44	3.35	4.60	2.51
66	**0.45**	8.84	8.54	10.99	6.95	1.37	1.46	2.10	0.98	3.35	3.38	4.64	2.55
67	**0.51**	10.33	10.20	12.21	8.21	1.56	1.70	2.26	1.06	3.79	3.91	5.00	2.75
68	**0.51**	10.37	10.10	12.37	8.21	1.56	1.68	2.28	1.08	3.78	3.88	5.05	2.81
69	**0.51**	10.36	10.02	12.46	8.07	1.55	1.67	2.31	1.06	3.76	3.84	5.11	2.74
70	**0.62**	13.16	12.75	15.29	10.77	1.92	2.09	2.76	1.33	4.66	4.82	6.13	3.43
71	**0.66**	13.84	13.62	16.16	11.37	2.00	2.24	2.91	1.40	4.85	5.16	6.47	3.59
72	**0.73**	15.44	14.79	17.61	12.82	2.23	2.42	3.16	1.59	5.44	5.59	7.04	4.09

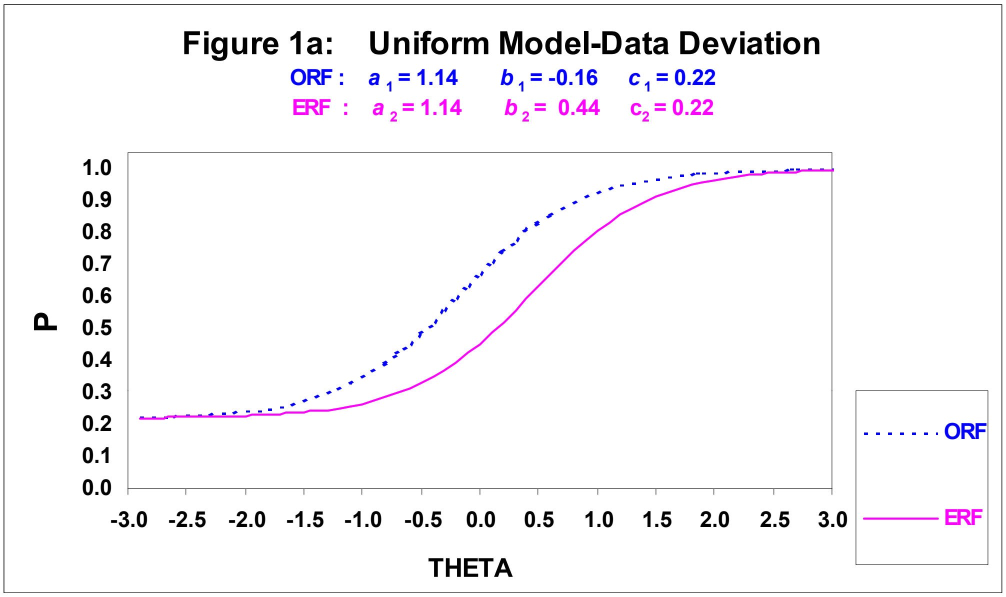


Figure 1a: Uniform Model-Data Deviation
ORF : a 1 = 1.14 b 1 = -0.16 c 1 = 0.22
ERF : a 2 = 1.14 b 2 = 0.44 c 2 = 0.22
P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
THETA
ORF
ERF

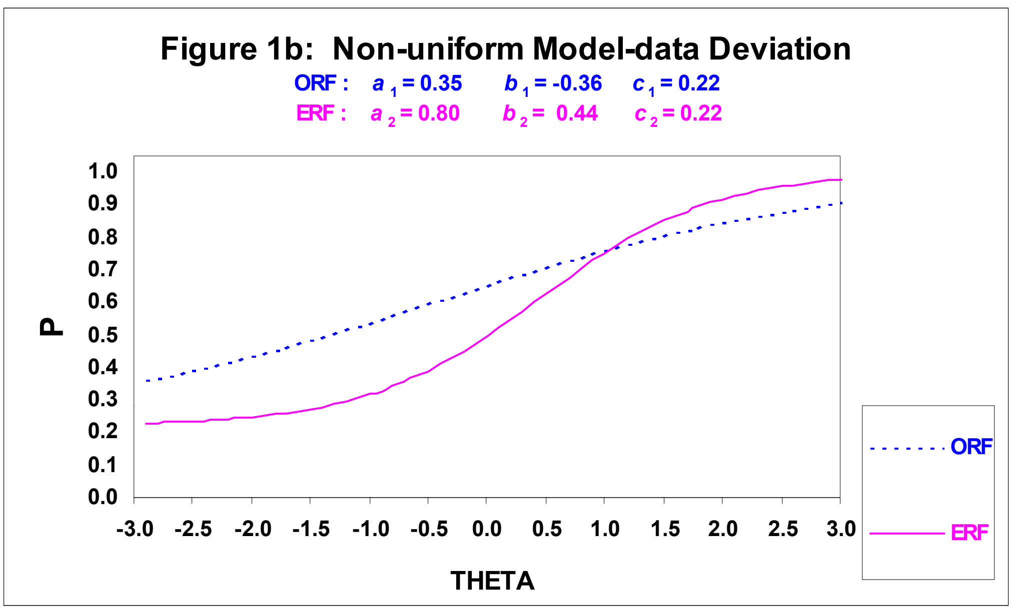


Figure 1b: Non-uniform Model-data Deviation
ORF : a 1 = 0.35 b 1 = -0.36 c 1 = 0.22
ERF : a 2 = 0.80 b 2 = 0.44 c 2 = 0.22
P
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
THETA
ORF
ERF

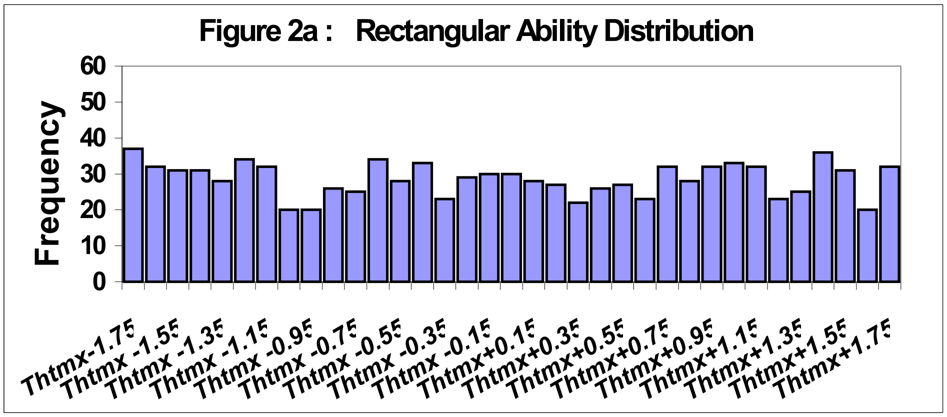


Figure 2a : Rectangular Ability Distribution
Frequency
0
10
20
30
40
50
60
Thtmx-1.75
Thtmx -1.55
Thtmx -1.35
Thtmx -1.15
Thtmx -0.95
Thtmx -0.75
Thtmx -0.55
Thtmx -0.35
Thtmx -0.15
Thtmx+0.15
Thtmx+0.35
Thtmx+0.55
Thtmx+0.75
Thtmx+0.95
Thtmx+1.15
Thtmx+1.35
Thtmx+1.55
Thtmx+1.75

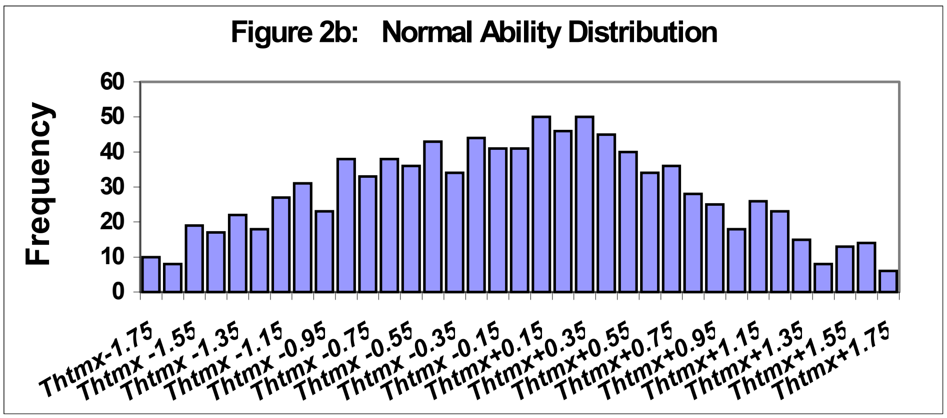


Figure 2b: Normal Ability Distribution
Frequency
0
10
20
30
40
50
60
Thtmx-1.75
Thtmx -1.55
Thtmx -1.35
Thtmx -1.15
Thtmx -0.95
Thtmx -0.75
Thtmx -0.55
Thtmx -0.35
Thtmx -0.15
Thtmx+0.15
Thtmx+0.35
Thtmx+0.55
Thtmx+0.75
Thtmx+0.95
Thtmx+1.15
Thtmx+1.35
Thtmx+1.55
Thtmx+1.75

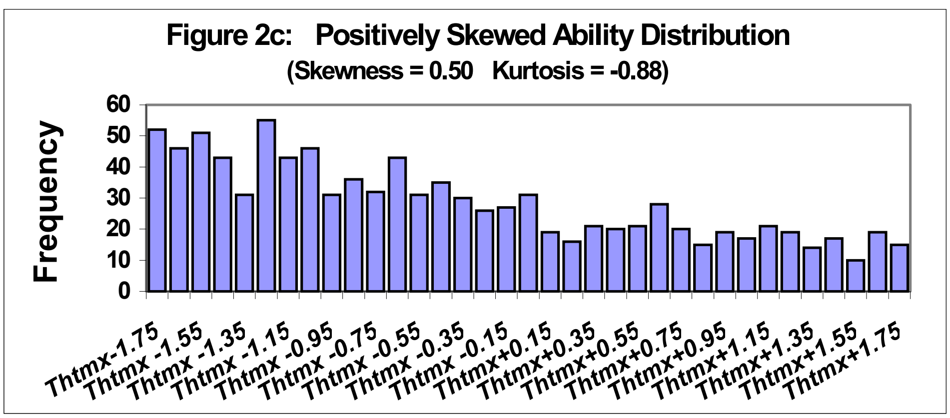


Figure 2c: Positively Skewed Ability Distribution
(Skewness = 0.50 Kurtosis = -0.88)
Frequency
0
10
20
30
40
50
60
Thtmx-1.75
Thtmx -1.55
Thtmx -1.35
Thtmx -1.15
Thtmx -0.95
Thtmx -0.75
Thtmx -0.55
Thtmx -0.35
Thtmx -0.15
Thtmx+0.15
Thtmx+0.35
Thtmx+0.55
Thtmx+0.75
Thtmx+0.95
Thtmx+1.15
Thtmx+1.35
Thtmx+1.55
Thtmx+1.75

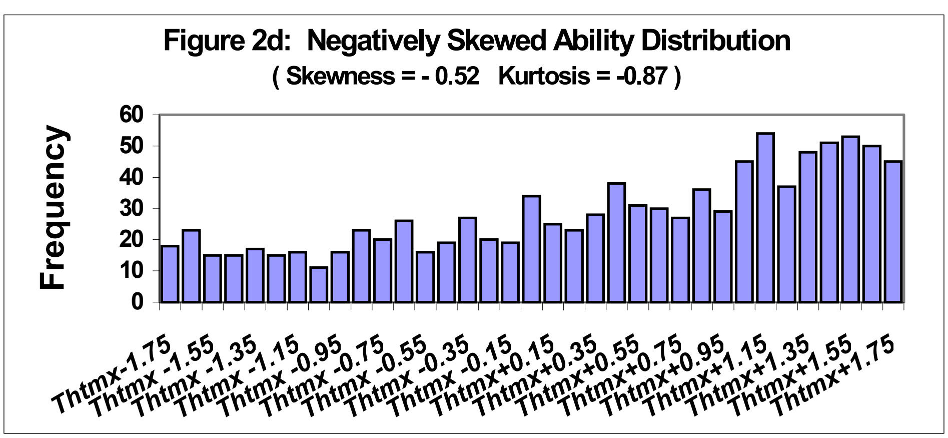


Figure 2d: Negatively Skewed Ability Distribution
(Skewness = - 0.52 Kurtosis = -0.87)
Frequency
0
10
20
30
40
50
60
Thtmx-1.75
Thtmx -1.55
Thtmx -1.35
Thtmx -1.15
Thtmx -0.95
Thtmx -0.75
Thtmx -0.55
Thtmx -0.35
Thtmx -0.15
Thtmx+0.15
Thtmx+0.35
Thtmx+0.55
Thtmx+0.75
Thtmx+0.95
Thtmx+1.15
Thtmx+1.35
Thtmx+1.55
Thtmx+1.75

Figure 3a: Performance of Zc on Uniform vs Non-uniform Deviations

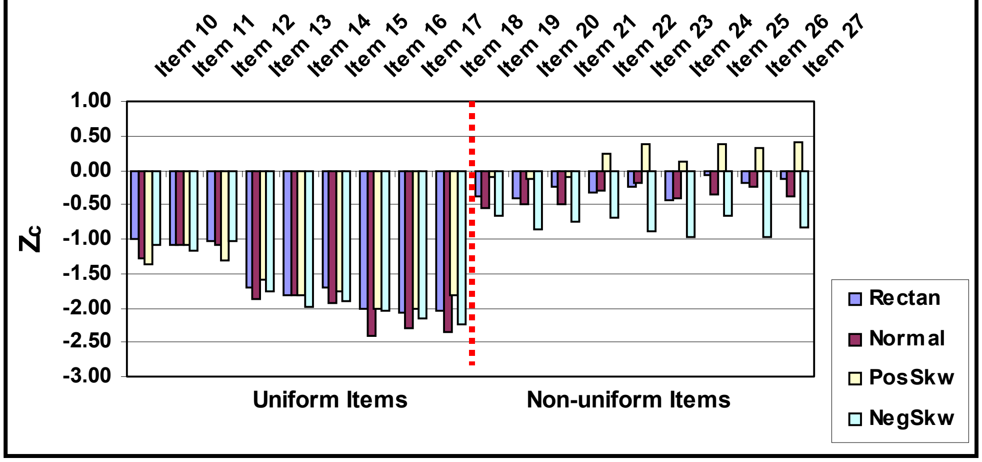


Figure 3b: Performance of Z2 on Uniform vs Non-uniform Deviations

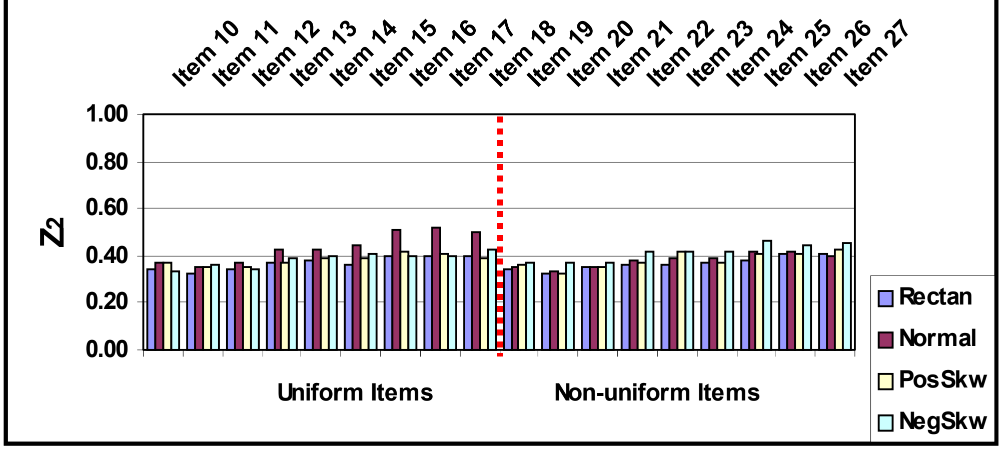


Figure 3c: Performance of Z3 on Uniform vs. Non-uniform Deviations

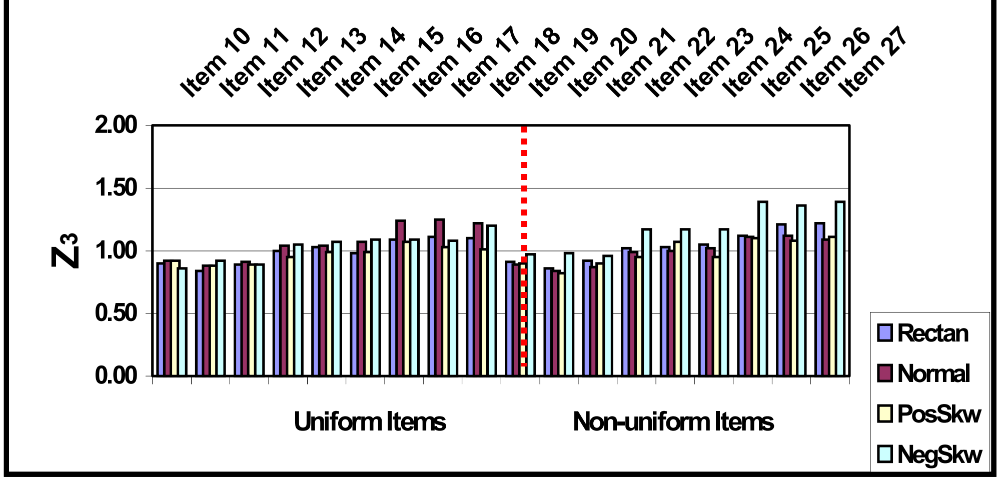


Figure 4a: Relationship Between Z2 and Degree of Deviation (N=400)

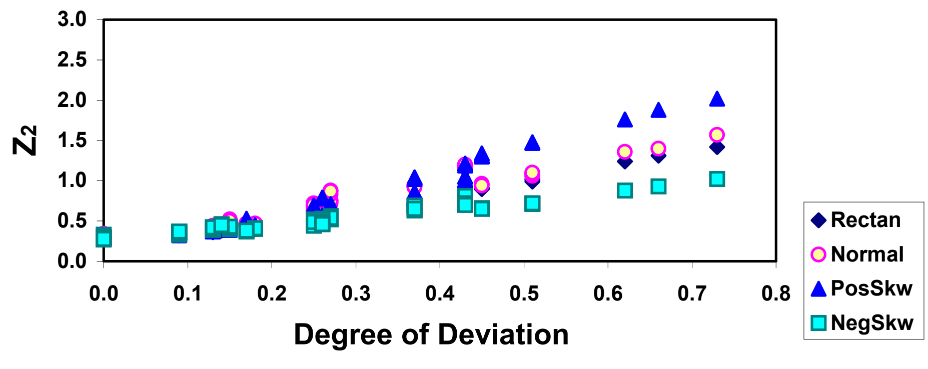


Figure 4b: Relationship Between Z3 and Degree of Deviation (N=400)

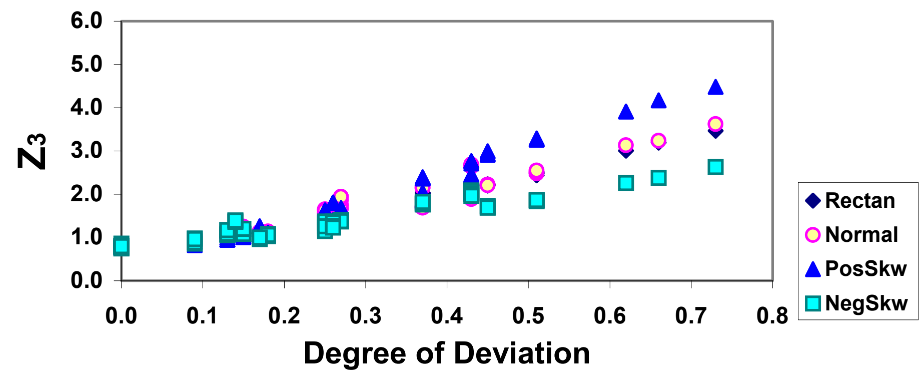


Figure 4c: Relative Stability of Z2 (Baseline=Normal; N=400)

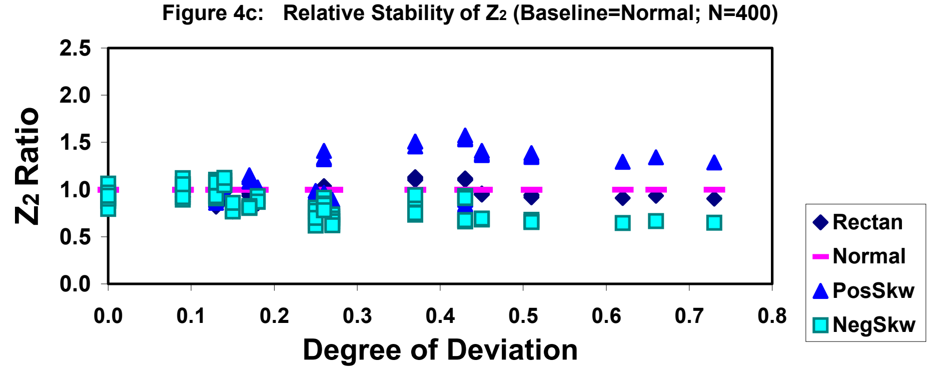
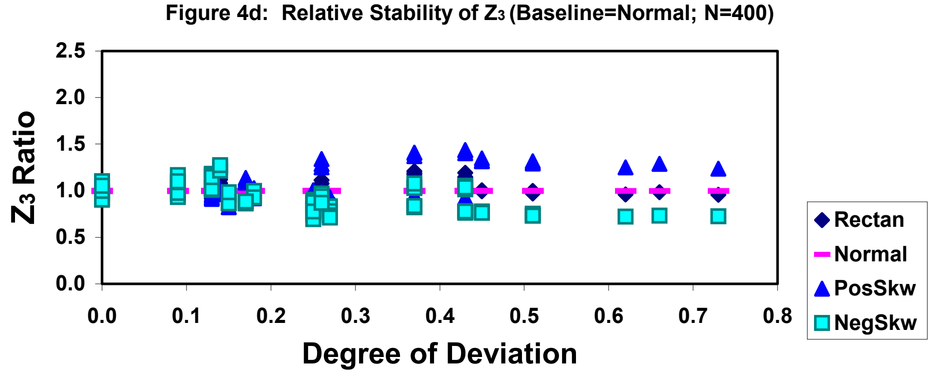


Figure 4d: Relative Stability of Z3 (Baseline=Normal; N=400)



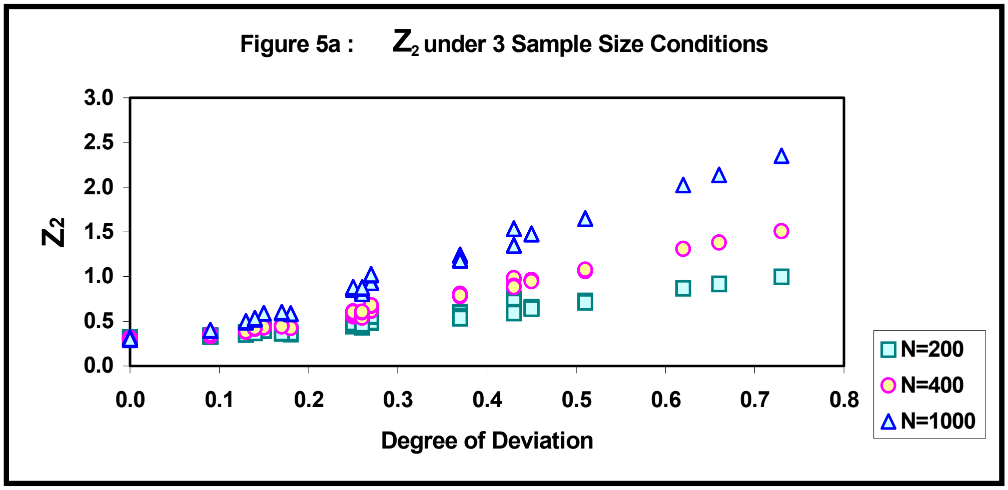


Figure 5a : Z_2 under 3 Sample Size Conditions
Z2
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Degree of Deviation
N=200
N=400
N=1000

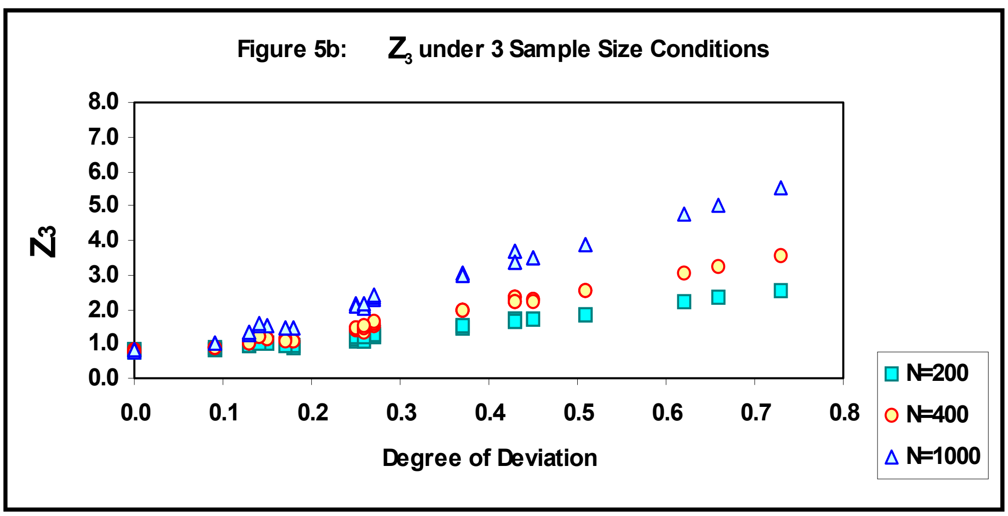


Figure 5b: Z_3 under 3 Sample Size Conditions
Z3
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Degree of Deviation
N=200
N=400
N=1000