Craig, S. B., & Harvey, R. J. (2004, April). *Using CAT to reduce administration time in 360° performance assessment.* In Craig, B. (Chair), *360, The next generation: Innovations in multisource performance assessment.* Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

Using CAT to Reduce Administration Time in 360° Performance Assessment

S. Bartholomew Craig
North Carolina State University

Robert J. Harvey
Virginia Tech

Among the challenges when collecting 360 performance assessment data are the need to (a) identify ratings that may be of questionable quality, and (b) minimize the amount of administration time. This study evaluated item response theory (IRT) *appropriateness* (or *person-fit*) indices and computer-adaptive testing (CAT) as strategies for addressing these issues. Using a sample of over 200,000 ratings collected using the the *Benchmarks*® 360° instrument (Center for Creative Leadership, 2000), exploratory factor analyses identified a 40-item maladaptive-behaviors scale, which we calibrated using Samejima's (1969) IRT graded response model (GRM). Results for 10,000 simulated CAT sessions showed that CAT can dramatically reduce 360 administration time, and that person-fit indices may be useful in identifying questionable response profiles.

Employees in modern organizations frequently complain of being "surveyed to death." Wide-scale adoption of many of the technologies advocated by I-O psychologists, such as climate surveys, needs assessments, and performance appraisals, has resulted in a data collection frenzy that arguably borders on the oppressive for many organizational members. The popularity of 360° performance assessment in particular has exacerbated this problem: whereas traditional performance appraisals use only one rater per ratee, 360° assessments may collect data from five or more raters per ratee, dramatically increasing the probability that a single rater will find him- or herself responsible for providing multiple ratings in any given rating period.

At best, this repeated surveying results in ratings being generated under suboptimal conditions, such as high time pressure and fatigue. At worst, forms may be completed haphazardly, or not at all. Given the importance of obtaining sound performance measures, it is critical that technologies be developed and implemented in organizations to reduce the amount of time required to collect performance-rating data without compromising data quality.

One technology that offers significant promise in this regard is computerized adaptive testing (CAT). CAT involves administering scale items via computer, with each item being dynamically selected from a large pool of calibrated items to be optimal for that given rater; each item in the pool has known psychometric characteristics that were previously quantified using item response theory (IRT). After the rater makes each response, the CAT administration software uses IRT methods to estimate the rater's latent construct score (based on the items presented up to that point), and then select the item that would be expected to be the most diagnostic at that particular level on the latent construct. The standard error (SE) of the score is also re-estimated after each item response, allowing a real-time index of the precision with which the rater's true standing on the latent construct is estimated. Testing may be terminated when the SE decreases to a predetermined level of precision, or when a predetermined maximum test length is achieved. Because CAT minimizes the presentation of inappropriate items, the number of items administered can often be reduced dramatically (typically, by up to half in traditional ability-type tests) while still providing precision comparable to that seen in the full-length test (e.g., Embretson & Reise, 2000).

Although CAT is a proven technology, having been adopted in a variety of high-stakes testing programs including the Scholastic Achievement Test and Graduate Record Exam, research is only beginning to assess its usefulness in organizational settings (e.g., Embretson & Reise, 2000; Harvey, 2003; Overton, Harms, Taylor, & Zickar, 1997), with applications of CAT to 360° representing an area in particular need of research. This study will provide a "proof of concept" by using actual rater data to assess the degree of reduction in instrument length via CAT that may be achieved using 360° assessments. We obtained a sample of over 200,000 ratings of managers by their coworkers (superiors, peers, subordinates) on the conventional (non-adaptive) version of the *Benchmarks*® 360° assessment instrument (Center for Creative Leadership, 2000) between 2000 and 2003, and used these ratings to simulate the CAT administration process (i.e.,

by dynamically selecting items as would occur in a real CAT administration, using raters' actual responses to those items; e.g., see Burnkrant & Harvey, 2000).

Using actual rater responses, the primary goal of this study was to determine determined which items — and how many items — would have been presented had the *Benchmarks*® 360° instrument been administered using CAT. Based on prior research on ability and personality tests (e.g., Brown & Harvey, 1998; Embretson & Reise, 2000; Keller & Harvey, 1999), we hypothesized that our simulation would indicate that substantial reductions in 360° assessment length can be achieved by using CAT. The only assumption required to conduct this procedure was that actual raters would provide the same ratings in a CAT situation as they gave when completing the identical items in the paper-and-pencil version. Although this is ultimately an empirical question, we felt comfortable in assuming that administration modality would not significantly alter raters' answers.

Samejima's (1969) graded-response model (GRM) was used to model the item-response process; the GRM offers many advantages over simpler number-right scoring methods for 360 instruments, given the popularity of ordered, Likert rating scales in such surveys. In the GRM, the probability that a response will fall in ordered category *k* (where *k* ranges from 0..*m* response alternatives) or higher can be defined as:

$$P_k^*(q) = \frac{1}{1+e^{-a(q-b_k)}} \qquad [7]$$

and in category $k + 1$ as

$$P_{k+1}^*(q) = \frac{1}{1+e^{-a(q-b_{k+1})}} \qquad [8]$$

where the limiting cases are defined by

$$P_{k=0}^*(q) = 1 \qquad [9]$$

and

$$P_{k=m+1}^*(q) = 0\,. \qquad [10]$$

Category response curves (CRCs) can be defined to quantify the probability of the respondent picking a response in a given category *x* (where $x = 0..m$) as:

$$P_{x=k}(q) = P_k^*(q) - P_{k+1}^*(q)\,. \qquad [11]$$

When IRT scoring is used, a*ppropriateness* (or person-fit) *indices* can be calculated for each person; these indices provide a measure of the internal consistency of the response profile (where in this context, "consistent" reflects the presence of responses to the items that correspond with the responses *expected* given the known IRT parameters for each item). For the GRM (e.g., see van Krimpen-Stoop & Meijer, 2002, p. 165-1666), person-fit indices based on residuals and the standardized height of the likelihood function are possible.

For indices based on the height of the likelihood function, Drasgow et al. (1985) defined a version of their Z_3 index for polytomous models in which the log-likelihood function is expressed as:

$$\ln(\mathrm{L}_0) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \sum_{j=0}^{m} d_j(x_i) \ln P_{ij}(q) \qquad [13]$$

where

$$d_j(x_i) = 1 \qquad [14]$$

if *j* equals the observed item response, and zero otherwise. The likelihood values need to be standardized, using

$$\mathrm{M} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \sum_{j=0}^{m} P_{ij}(q) \ln P_{ij}(q) \qquad [15]$$

and

$$\mathrm{S}^2 = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left[\sum_{j=0}^{m} \sum_{\mathrm{h}=0}^{\mathrm{m}} P_{ij}(q) P_{ih}(q) \ln P_{ij}(q) \ln\left(\frac{P_{ij}(q)}{P_{ih}(q)} \right) \right]. \qquad [16]$$

to produce the standardized likelihood as

$$\mathrm{Z}_3 = \frac{\mathrm{L}_0 - \mathrm{M}(\hat{q})}{\sqrt{\mathrm{S}(\hat{q})}} \qquad [5]$$

Because the numerator of Z_3 reflects the mean-deviation maximum likelihood estimate, higher values of this index denote more appropriate profiles.

In contrast, residual-based person-fit indices define lack of fit based on the summed differences between expected and observed item responses. In the GRM, the expected response for a given item can be defined as:

$$E(X_i \mid q) = \sum_{j=0}^{m} j P_{ij}(q) \qquad [12]$$

where $P_{ij}(\theta)$ denotes the probability of scoring $X_i = j$ on item *i* conditional on θ (i.e., $P_{x=k}$). Wright and Masters (1982) developed a residual-based person-fit index; as discussed by van Krimpen-Stoop and Meijer (2002),

$$v = \frac{\sum_{i=1}^{N} \left[X_i - \sum_{j=0}^{m} j P_{ij}(q) \right]^2}{\sum_{i=1}^{N} \left[\sum_{j=0}^{m} (j - E(X_i \mid q))^2 P_{ij}(q) \right]} \qquad [17]$$

Both Z_3 and v should be sensitive to detecting aberrant response profiles composed of ratings that seem inconsistent with the remaining ratings for the scale (e.g., as might occur from a rater who is hurrying through the survey and not closely reading the items), and as such may offer promise in a 360 context via offering test administrators a tool that can be used to spot profiles that may deserve closer examination. Thus, in addition to our primary goal of assessing the degree to which CAT could be used to reduce the number of 360 items administered, a secondary objective of this study was to evaluate the performance of IRT person-fit indices using two approaches. First, we sought to determine the relative number of real-data raters who would have been flagged as potentially aberrant had person-fit indices been used. Given that the raters who completed the *Benchmarks*® 360° assessment did it under "real" conditions, we hypothesized that most ratings would score well in terms of these rating-quality indices.

Second, we examined the degree to which such indices might be useful in CAT as indicators of situations in which a shortened test might *not* provide a good index of the person's score (i.e., in comparison to the one that would have been obtained had the full-length survey been given). That is, the primary goal of CAT is to reduce testing time as much as possible while still producing a score for each scale that is highly similar to the score that would be obtained from the full-length test. However, especially for rating targets who are not extreme in their true scores on the scale, it is possible that in a much shorter version of the test CAT might produce θ score estimates that diverge significantly from the θs based on the full-length survey. For example, in a situation in which the 360 dimension defines *undesirable* leader behaviors, it might be possible for a manager who only performs a few highly undesirable behaviors to, in a CAT, look much better than is actually the case (i.e., if the few items on which he/she is highly inappropriate were not included among the items presented in the CAT).

To address this question, we examined the degree to which IRT person-fit indices computed in shortened versions of the test would be predictive of the magnitude of the *difference* that was present between the short-length θ score versus the θ score produced for the full-length test. Although this approach obviously would not be effective if *none* of the items on which the target receives deviant ratings are actually presented in the short-length CAT item pool, if any of the items on which the manager receives bad ratings in the full pool make it into the CAT pool, we hypothesized that the appropriateness indices would provide at least some warning of this situation. In actual CAT administrations, values of the IRT person-fit indices could be monitored by the CAT program as the survey is being given, and if troublesome values are found, larger numbers of items (up to and including the full-length test) could be given to obtain the most accurate θ estimate possible.

Method

Participants and Measures

A sample of N = 249,464 individuals who completed the *Benchmarks*® 360° instrument (Center for Creative Leadership, 2000) between 2000 and 2003 was provided by the Center for Creative Leadership. This instrument contained two main sections of interest in this study: Section 1 contained 115 items dealing with a range of leadership and supervisory behaviors rated on a frequency-of-observation scale ranging from 1 = *not at all* through 5 = *to a very great extent* (higher scores typically define more adaptive behaviors), whereas Section 2 contained 40 items defining maladaptive leader behaviors rated on a scale from 1 = *strongly disagree* through 5 = *strongly agree* that the item is characteristic of the ratee (here, higher scores reflect undesirable performance).

Analyses

Exploratory factor analyses using the common factor model (with squared-multiple correlation estimates of communalities) were used to guide the identification of a suitable 360 item pool for IRT scoring and CAT presentation. In particular, we sought to identify a relatively large pool of items that was dominated by a single underlying factor. Once the pool was identified, the item responses were calibrated using the GRM, the 1-5 Frequency responses were analyzed using MULTILOG (version 6.3); for person-fit indices, the polytomous Z_3 was computed using Equations 12-16, and v was computed using Equation 17. Given the large number of raters available, raters who produce any missing responses were excluded, producing a sample of N = 206,217. For the MULTILOG analyses, for logistical reasons a random sample of N = 3,000 was drawn from that sample; however, scores on the full N = 206,217 sample were estimated once the items were calibrated (Bayes EAP scoring using a Normal prior was used to estimate θ).

When conducting the simulated CAT administrations, the CAT was begun by sampling three items at random from the full pool, estimating the initial θ, and then selecting all subsequent items based on identifying the item with the highest amount of information at the provisional θ estimate. Estimates of θ and the person-fit indices were computed after selecting each item, and the CAT process continued until all items were administered. Primary comparisons were those between 10-, 20-, and 30-item CATs versus the full 40-item pool. Sensitivity to detecting inappropriate profiles via the various appropriateness indices was assessed by correlating them with absolute-value difference scores (for a given number of CAT items administered) between the current θ estimate and the θ estimate using the full scale. Suitability of CAT for this 360 application was assessed by comparing θ estimates computed for varying length tests against the full-pool θ estimate.

Results

Eigenvalues from the exploratory common factor analyses of the 155-item pool formed from the Section 1 and 2 items are shown in Figure 1. This scree plot strongly suggested that although the *Benchmarks*® 360° instrument may be scored to produce a large number of subscales for feedback purposes, at

a big-picture level this item pool defined perhaps three or four major domains. Examinations of rotated low-dimensionality oblique factor solutions (using Harris-Kaiser orthoblique rotation) revealed factors that largely defined the time-honored Ohio State LBDQ dimensions seen in leadership research (i.e., Consideration and Initiating Structure), plus a third major factor dealing with the maladaptive behavioral items described in the second section.

Accordingly, given our desire to identify an item pool composed of a relatively sizeable number of items dominated by a single underlying major dimension in the IRT analyses, we focused on the 40-item pool of items describing maladaptive managerial behaviors. As the scree plot (see Figure 2) of eigenvalues makes clear, this 40-item pool is indeed dominated by a single underlying dimension (using squared-multiple correlation estimates of communalities, the first factor explained 84% of the total common variance). Given prior research (e.g., Drasgow & Parsons, 1983), these results were viewed as indicating that this item pool is sufficiently unidimensional to be calibrated using IRT.

Results of the GRM calibrations are presented in Table 1, as are classical item-total correlations and the item loadings on the first unrotated common factor. An inspection of the results shown in Table 1 shows that this 40-item scale exhibits strong item-totals and factor loadings, as well as generally strong *a* (discrimination) IRT parameters. An inspection of the full-scale test information function (TIF; see Figure 3) and test standard error (TSE) function (see Figure 4) reveals similarly highly desirable findings; in both cases, these functions depict a strong item pool that offers very good information (and corresponding low SE) across a wide range of θ scores (e.g., the TSE is below 0.20, and essentially linear, from –0.5 θ all the way through +3.0 θ). Given that higher scores on this scale are indicative of questionable managerial effectiveness, the fact that this item pool provides good precision from θ = -1.0 through θ values above +3.5 is highly desirable (i.e., the highest precision lies in the range of θ score values that likely represent the largest causes for concern regarding managerial leadership-development needs).

In terms of the question of CAT, Table 2 presents correlations among the θ and person-fit indices for CAT administrations of varying length (i.e., 10-, 20, 30-, and full 40-item administrations); Figure 5 presents frequency distributions of θ for varying test length; Figure 6 compares θ estimates for 10- and 20-item subsets against full-length θ; and Figure 7 presents scatterplots comparing person-fit indices computed using the 10-item CAT against full-pool θ as well as the difference between full-pool versus 10-item θ. Given that a small subset of the 10,000 raters produced perfect-low scores (i.e., giving the minimum rating of '1' – or highly effective – to all 40 items), an N = 9,420 sample was formed by excluding raters who produced perfect-low profiles. That is, because an IRT-based person-fit index will invariably indicate that a person who produced a perfect-low score is "appropriate" – as well as the fact that it would be impossible for sampling-based issues to lead to underrepresentation of ratings that were *inconsistent* with the rest of the profile – person-fit results for perfect-negative profiles were not of interest for analyses designed to determine the degree to which person-fit indices are predictive of item-sampling-based problems in CAT.

In terms of the question of whether test-length can be reduced appreciably via CAT, these results indicate that substantial reductions can easily be achieved without compromising measurement precision. For example, the results in Table 2 show $r > .96$ values comparing θ estimated using the 10-item CAT (i.e., one-quarter the full item pool) versus both the number-right and θ estimate computed using all 40 items. Although correlations in excess of .96 are certainly large, they do not reflect scores that are *completely* interchangeable, however. For example, although Figure 5 reveals that overall score distributions are quite similar in CAT subsets of widely varying length, Figure 6 shows that there is indeed some degree of scatter when comparing θ estimates computed using CAT versus full item pools.

Not surprisingly, given the maximum-information method for selecting the next CAT item, the Table 1 results indicate that some items are indeed presented at appreciably higher rates than others. However, in the typical 360° assessment situation raters, test-security and item-exposure issues are not likely to be troublesome (unlike the situation in high-stakes ability testing, in which item security and avoiding item over-exposure are critical), particularly given the fact that many raters may complete the same survey numerous times in the context of rating different targets (or when rating the same target repeatedly over time). As hypothesized, the largest degree of subset versus total-pool θ score divergence tends to fall at the lower end of the θ scale (see Figure 6), a finding that presumably reflects the fact that some rating targets who score generally well in terms of these maladaptive leader behaviors are rated poorly on a subset of items (and that this fact is not being detected as effectively in shorter CAT administrations, given that the items on which the bad ratings occur are not always presented in CAT).

Regarding the IRT person-fit indices, an inspection of the results in the upper portion of Figure 7 shows that, as expected, a relatively small number of raters exhibit values that fall toward the clearly-inappropriate range of Z_3 (lower scores, especially below approximately –2) and v (higher scores). With respect to the ability of these person-fit indices to detect situations in which appreciable differences exist between CAT-based versus full-item-pool θ score estimates, correlations between the absolute-value difference between 10-item CAT versus full-pool θ and each of the person-fit indices were r = 0.17, 0.23, 0.21, -0.20, -0.31, and -0.30 for v-10, v-20, v-40, Z_3-10, Z_3-20, and Z_3-40, respectively (all values $p < .0001$ in the N = 9,420 sample). Although all correlations were significant and in the expected direction, these results clearly indicate that IRT person-fit indices cannot flag all cases in which appreciable differences exist between CAT versus full-length θ estimates. However, such indices may still be helpful with respect to this goal.

Similar conclusions are indicated from the results presented in Figure 7, which show that the differences between the most divergent θ scores are not especially large overall (i.e., for the 10-item CAT versus full pool θ, the median absolute difference is only 0.152, with interquartile range of only 0.197; this relatively small average difference,

which is smaller than the standard error seen for most of the effective range of the scale, also argues in favor of the above conclusion regarding the success of CAT in maintaining precision in a dramatically shortened test). Although there is some tendency for raters having higher absolute differences in θ to also exhibit more troublesome values of the person-fit indices (i.e., lower values of Z_3, and higher values of *v*, are indicative of poorer fit), this association is hardly overwhelming; indeed, at least in this very short 10-item CAT, a number of people who show very little difference between their full versus CAT θ estimates show sizable values for the person-fit indices.

Discussion

Regarding our primary goal – i.e., to determine the degree to which overall test length can be reduced via CAT without unduly reducing the convergence between θ scores estimated using CAT versus the full item pool – the results presented above indicate that CAT offers significant promise in the context of 360° surveys that are constructed in the fashion we followed in this study (i.e., with relatively large item pools that are dominated by a general underlying factor). In terms of the secondary question of whether person-fit indices can aid in identifying cases in which longer administrations (if not the full item pool) should be given, the above results suggest that person-fit indices may play a potentially useful role. Additional research on this topic is needed, particularly studies that examine the degree to which such indices are able to detect aberrant profiles that exhibit varying degrees and types of known-bad responses (e.g., Harvey, 2004).

Of course, the application of CAT to 360° assessment does pose some challenges, primarily with respect to the issue of producing developmental feedback reports. For example, because different raters who judge the same target may be presented with minimally overlapping sets of items, the traditional practice of presenting ratees with mean item-level results broken down by rating source may need to be revisited. One possible method to address this issue would be to ensure that the CAT began using a common subset of initial items on which all targets would be rated; after evaluating these answers, the remaining CAT items could be selected so as to be maximally informative for that particular target. Thus, at least some common-metric of shared item-level feedback would be possible when developing feedback reports for all targets.

Additionally, although in theory measurement precision is maximized when latent construct scores are estimated via IRT-based methods (e.g., maximum likelihood estimation) relative to simpler number-right methods, such scores may suffer from the limitation of being difficult to explain to feedback recipients who lack statistical training. One possible solution to this issue (as well as the common-item issue) may lie in the use of the "homogenous item cluster" concept seen in assessment instruments like the *Hogan Personality Inventory* (Hogan Assessment Systems, 1995).

References

Brown, R. D., & Harvey, R. J. (1998, April). *Computer-adaptive testing and test-retest reliability in a "Big-Five" personality inventory.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas.

Burnkrant, S, & Harvey, R. J. (2000, April). *Establishing baserates for the Z3 and F2 inappropriateness indices*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans.

Drasgow, R., & Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology, 72*, 10-18.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardization indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Drasgow, F., & Parsons, C. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

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

Harvey, R. J. (2003, April). *Applicability of binary IRT models to job analysis data.* In Meade, A. (Chair), *Applications of IRT for Measurement in Organizations.* Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando.

Hogan Assessment Systems (1995). *Hogan personality inventory.* Tulsa, OK: Author.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Keller, K. D., & Harvey, R. J. (1999, April). *Computer-adaptive testing with the MBTI.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161-176.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evauating person fit. *Applied Psychological Measurement, 25*, 107-135.

Overton, R.C., Harms, H.J., Taylor, L.R., & Zickar, M.J. (1997). Adapting to adaptive testing. *Personnel Psychology, 50*, 171-185.

Rogers, H. J., & Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47-57.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26*, 164-180.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

Table 1. *Item Parameter Estimates*

Item	load	r_{tot}	A	B1	B2	B3	B4	AvgSeq	Pct
1. Not manage different dept.	0.618	0.63	1.64	0.08190	1.400	2.10	3.23	3.3195	8.20
2. Neglects to focus on high-profile	0.589	0.61	1.50	-0.20600	1.480	2.15	3.66	8.4491	15.31
3. Promotion cause past competence	0.647	0.66	1.69	-0.26300	0.835	1.50	2.58	10.1282	21.02
4. Doesn't complete work	0.630	0.64	1.84	0.21200	1.680	2.32	3.41	2.0925	7.68
5. Not ready for more resp.	0.691	0.70	1.94	-0.16800	0.958	1.57	2.66	12.9677	28.91
6. No learn from mistakes	0.734	0.74	2.42	-0.11100	1.380	2.08	3.10	12.7322	95.12
7. Can't deal with participative boss	0.647	0.66	1.82	-0.08310	1.420	2.43	3.65	11.1598	16.67
8. Not adapt to culture	0.620	0.63	1.75	0.40100	1.850	2.43	3.59	2.1516	7.49
9. Arrogant	0.672	0.68	2.03	0.49800	1.530	1.98	2.99	10.8787	14.88
10. No understanding of other depts.	0.587	0.60	1.53	-0.01120	1.750	2.42	3.99	5.1308	10.06
11. Not adaptable	0.697	0.71	2.18	-0.00821	1.540	2.13	3.25	16.2236	47.99
12. Resists outside input	0.719	0.73	2.28	-0.37500	1.250	1.84	3.21	11.7503	66.48
13. Can't handle outside function	0.734	0.74	2.33	0.04580	1.250	1.93	2.90	12.3593	83.37
14. Unprofessional disagreements	0.621	0.63	1.88	0.49300	1.990	2.66	3.67	2.1774	7.71
15. Dictatorial	0.680	0.69	1.85	-0.03130	1.150	1.64	2.86	6.0988	10.21
16. Boss conflicts	0.515	0.53	1.37	0.48400	1.780	2.64	3.83	2.0505	7.35
17. Demeans subordinates	0.698	0.70	2.25	0.38200	1.500	2.02	3.01	13.7329	59.90
18. Doesn't use feedback to improve	0.766	0.77	2.55	-0.47300	0.968	1.69	2.81	7.6388	99.89
19. Won't share decision-making	0.710	0.72	2.19	-0.43700	1.100	1.65	2.87	12.8607	67.74
20. Won't resolve conflict among subords	0.678	0.69	1.99	-0.61000	0.946	1.79	3.03	8.4772	34.39
21. Leaves trail	0.717	0.73	2.26	0.26100	1.260	1.80	2.65	11.8286	64.60
22. Complexity overwhelms	0.604	0.61	1.68	0.11900	1.800	2.51	3.80	2.2635	7.65
23. Hires only on tech skills	0.610	0.62	1.70	-0.32400	1.280	2.49	3.74	10.2362	24.30
24. Emotionally volatile	0.637	0.65	1.85	0.40800	1.690	2.20	3.25	2.1592	7.86
25. No motivating for team effort	0.754	0.76	2.52	-0.25500	1.170	1.74	2.83	9.6911	99.92
26. Not adapted to culture	0.696	0.70	2.14	0.09690	1.520	2.13	3.23	15.3292	33.02
27. Bullies under stress	0.697	0.70	2.08	0.14400	1.200	1.67	2.80	14.6268	37.66
28. Hires narrow workers	0.697	0.71	2.23	-0.27400	1.110	1.99	3.13	14.4765	99.73
29. Bad under pressure	0.699	0.71	2.11	-0.18600	1.410	2.07	3.28	12.9668	38.09
30. Past competence	0.718	0.72	2.38	0.21500	1.370	2.11	2.93	10.5052	69.57
31. Selects won't work well	0.668	0.67	2.13	-0.16000	1.460	2.56	3.65	14.2468	42.96
32. Closed minded	0.717	0.73	2.19	-0.64000	0.844	1.43	2.62	11.7015	79.35
33. Overestimates ability	0.774	0.78	2.74	-0.22800	1.120	1.78	2.75	7.1165	98.57
34. Not team-builder	0.790	0.79	2.85	-0.28600	0.996	1.66	2.60	5.0582	98.83
35. Hard to keep up with current expect.	0.733	0.74	2.48	0.17100	1.430	2.02	2.97	9.3781	76.28
36. Self-promoter	0.726	0.73	2.51	0.25400	1.490	2.08	2.96	8.5226	72.55
37. Stuck as technical manager	0.704	0.71	2.15	-0.09610	1.100	2.00	3.04	15.9275	87.16
38. Lacks big picture	0.713	0.72	2.31	-0.34700	1.260	2.02	3.27	12.0783	75.79
39. Bosses rather than working with	0.765	0.77	2.77	0.10600	1.300	1.82	2.81	6.1275	83.57
40. Not encouraging	0.779	0.78	2.85	-0.09390	1.230	1.76	2.80	5.9356	91.94

Note. Item parameter estimates are computed using MULTILOG on N = 50,000 sample. Item-total correlations computed for total N = 206,217 sample. AvgSeq = average sequential position for the item in the N = 10,000 simulated CAT administrations. Pct = percentage of times in simulated CAT sessions that the item was presented in the 20-item test. Load = loading on unrotated first factor underlying the item pool.

Table 2. *Correlations Among q Estimates and Person-Fit Indices in CATs of I0-, 20-, 30- and Full Length*

	Mean	theta10	theta20	theta30	theta40
Mean	1.00000				
theta10	0.92623	1.00000			
	<.0001				
theta20	0.95012	0.97791	1.00000		
	<.0001	<.0001			
theta30	0.95881	0.96945	0.99284	1.00000	
	<.0001	<.0001	<.0001		
theta40	0.96349	0.96420	0.98819	0.99617	1.00000
	<.0001	<.0001	<.0001	<.0001	
se10	-0.57728	-0.77608	-0.74748	-0.73904	-0.73640
	<.0001	<.0001	<.0001	<.0001	<.0001
se20	-0.56386	-0.72726	-0.75016	-0.74141	-0.73891
	<.0001	<.0001	<.0001	<.0001	<.0001
se30	-0.55998	-0.71090	-0.73343	-0.74278	-0.74134
	<.0001	<.0001	<.0001	<.0001	<.0001
se40	-0.55565	-0.70081	-0.72276	-0.73296	-0.73984
	<.0001	<.0001	<.0001	<.0001	<.0001
v10	0.09540	0.06221	0.06256	0.06445	0.06135
	<.0001	<.0001	<.0001	<.0001	<.0001
v20	0.03821	-0.03905	-0.02235	-0.02073	-0.02480
	0.0002	0.0002	0.0301	0.0442	0.0161
v30	0.02446	-0.06343	-0.05539	-0.04586	-0.04935
	0.0176	<.0001	<.0001	<.0001	<.0001
v40	0.02784	-0.06848	-0.06094	-0.05220	-0.05134
	0.0069	<.0001	<.0001	<.0001	<.0001
z3_10	-0.09218	-0.05311	-0.05280	-0.05316	-0.04795
	<.0001	<.0001	<.0001	<.0001	<.0001
z3_20	-0.08073	-0.01140	-0.02747	-0.02844	-0.02083
	<.0001	0.2686	0.0077	0.0058	0.0432
z3_30	-0.08247	-0.00405	-0.01557	-0.02487	-0.01684
	<.0001	0.6945	0.1309	0.0158	0.1022
z3_40	-0.09615	-0.00675	-0.01688	-0.02578	-0.02441
	<.0001	0.5127	0.1013	0.0123	0.0178

Note. Values under correlation represent *p* value.

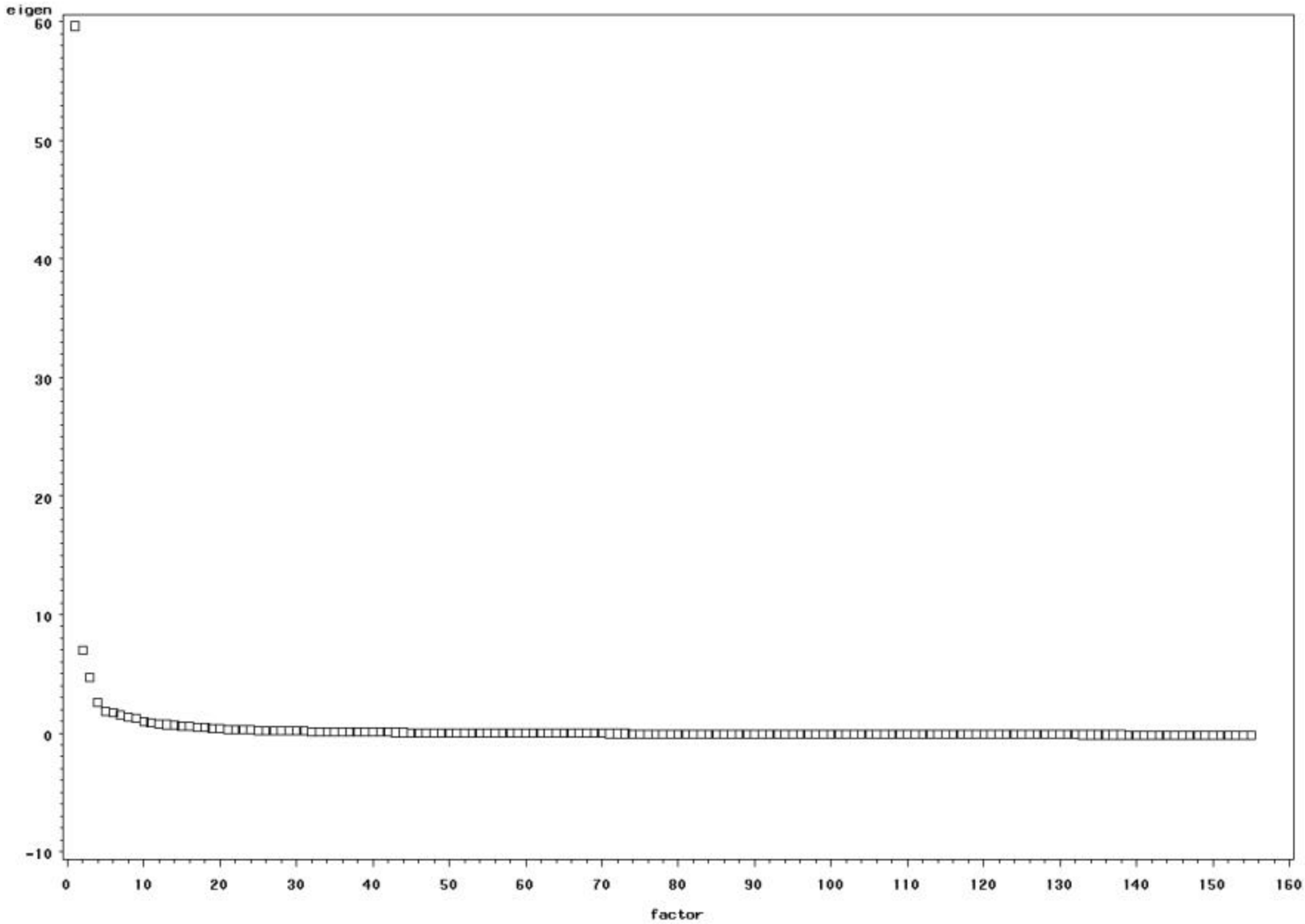

Figure 1. Scree plot of SMC-priors eigenvalues for *N* = 153,731 sample of raters having no missing responses on items from Section 1 and 2.

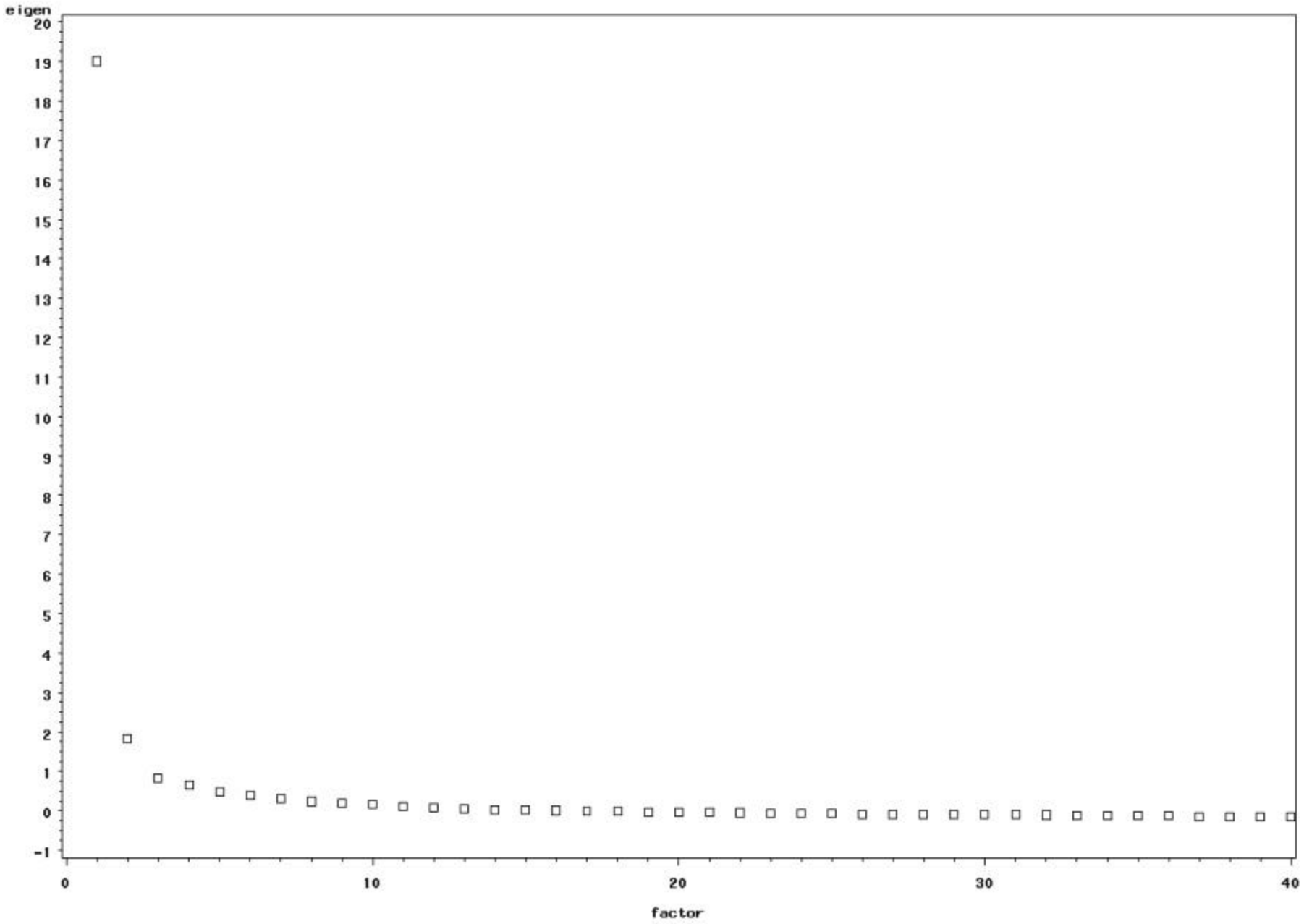

Figure 2. *Figure 1*. Scree plot of SMC-priors eigenvalues for *N* = 206,217 sample of raters having no missing responses on items from Section 2.

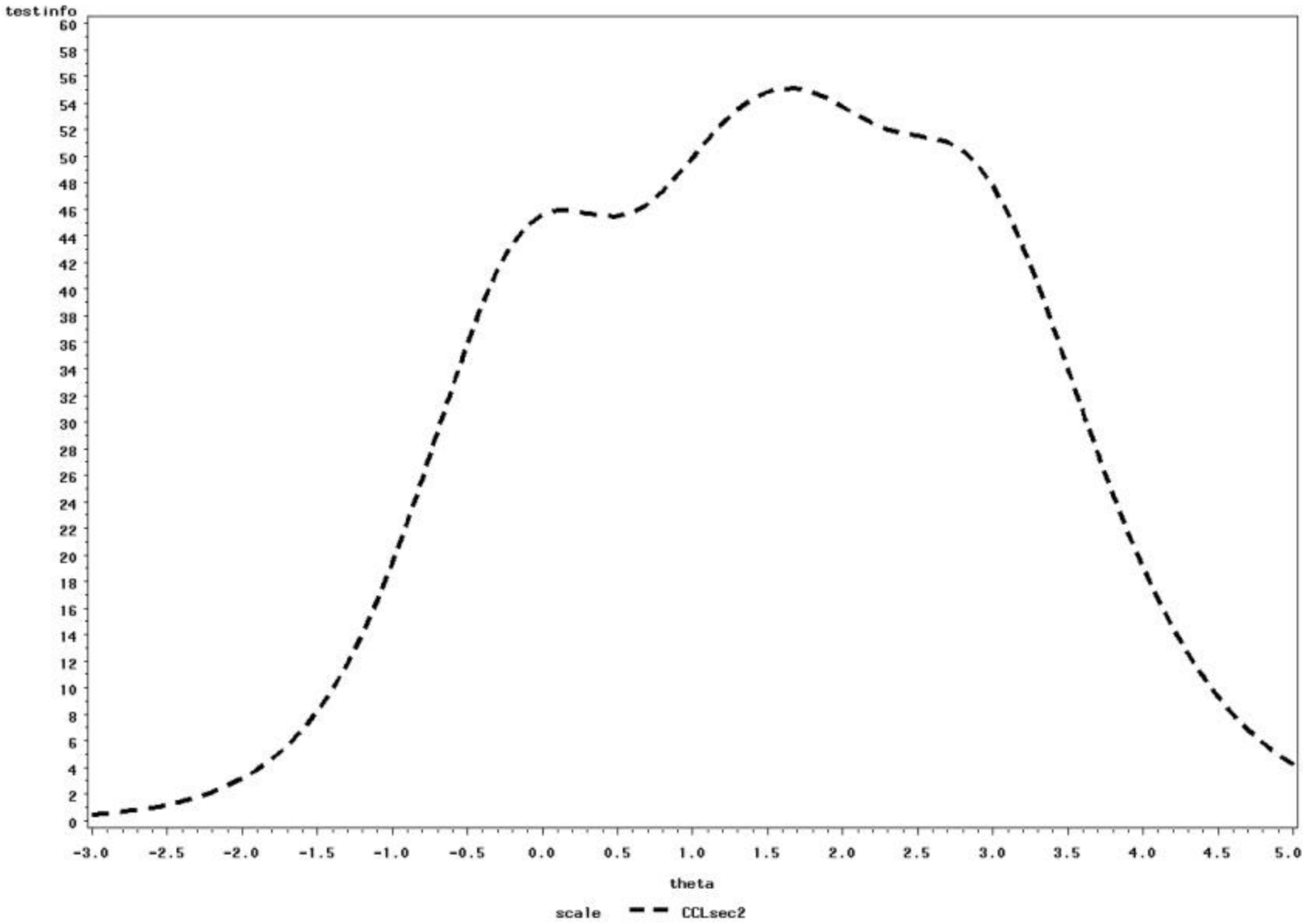

Figure 3. Test information function for N = 50,000 calibration sample.

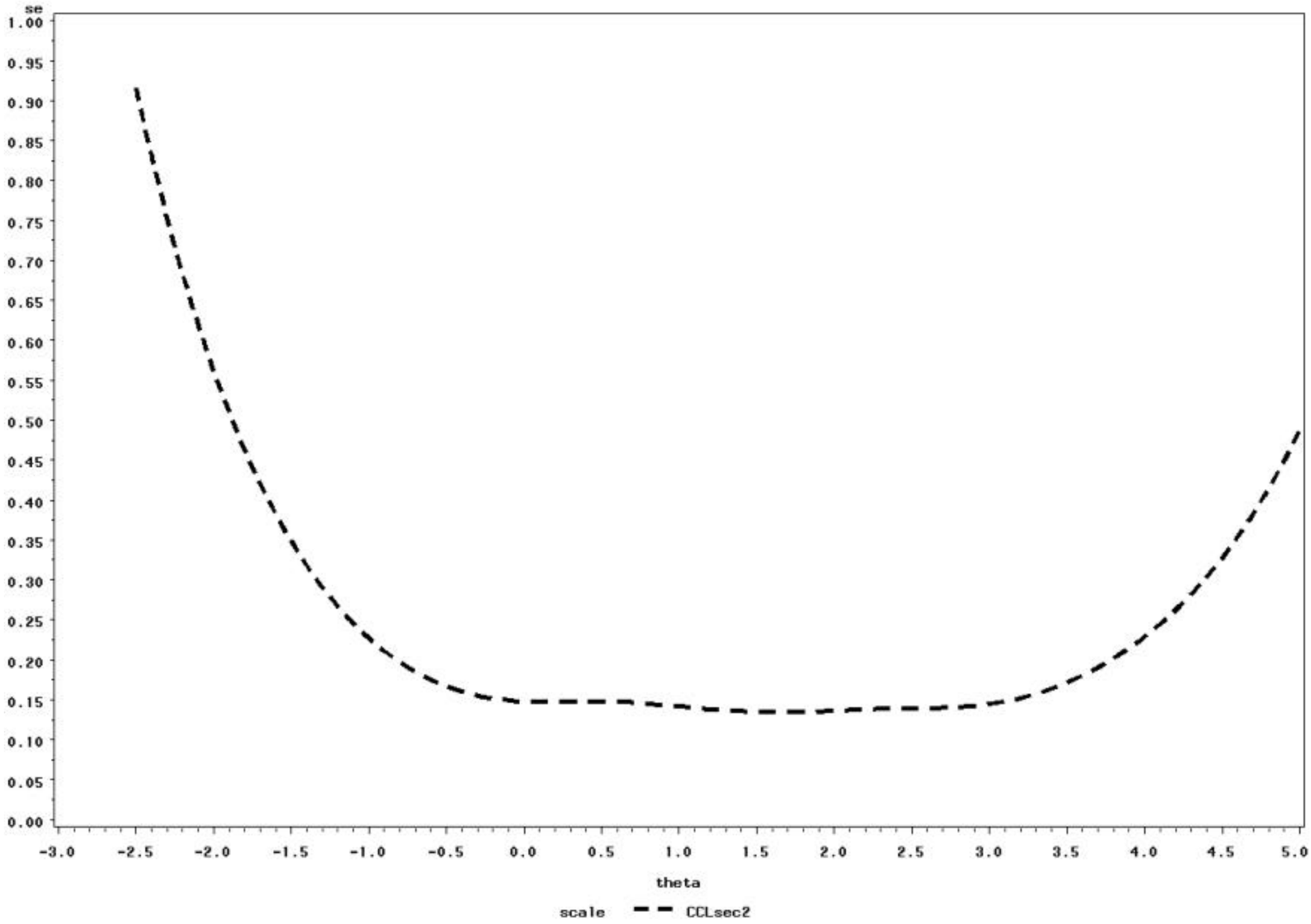

Figure 4. Test standard error function for N = 50,000 calibration sample.

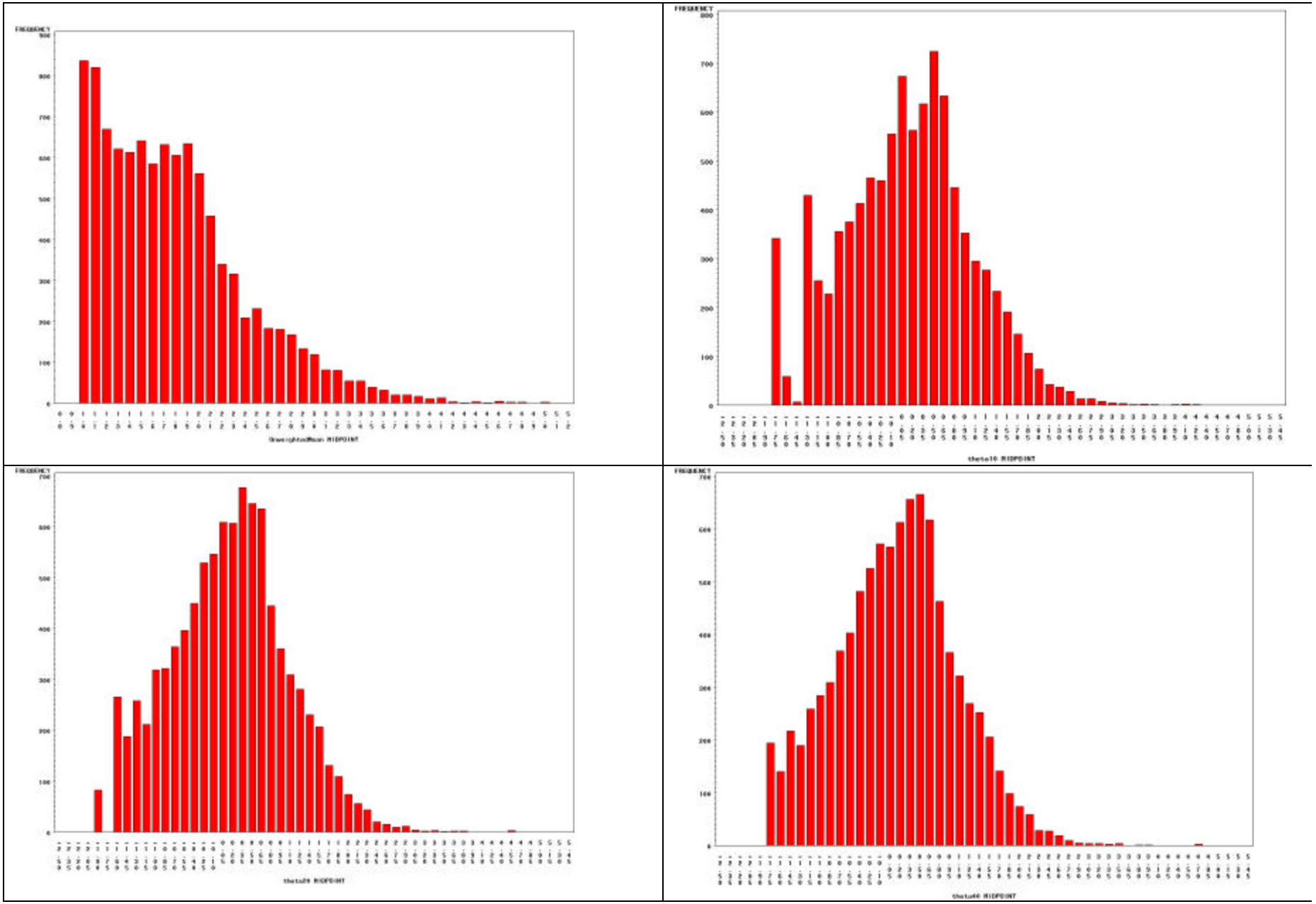

Figure 5. Distribution of N = 10,000 total scores (upper left) and EAP theta scores (N = 9,420 after excluding perfect low scores) for 10-, 20-, and 40-item pools (upper right, lower left, and lower right, respectively).

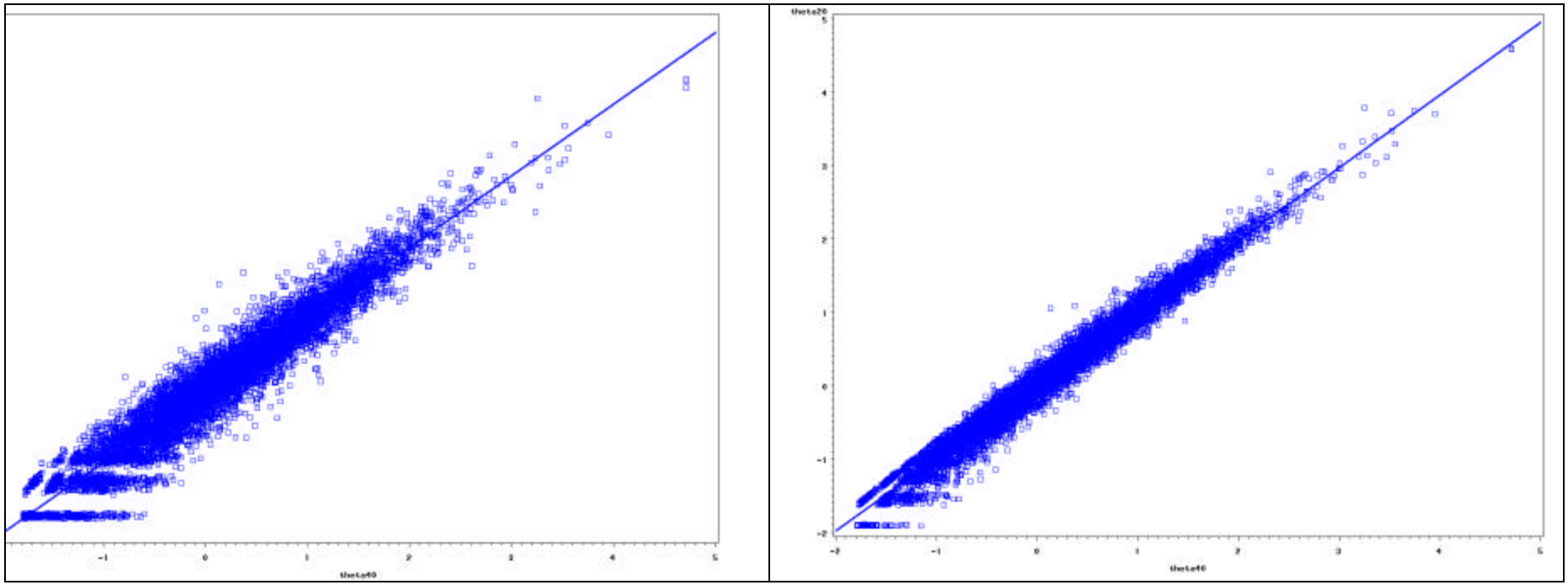

Figure 6. Scatterplots of theta scores computed for CAT administrations with y-axis representing estimates produced using 10- (left) and 20-item subsets (right); x-axis is score from full 40-item pool.

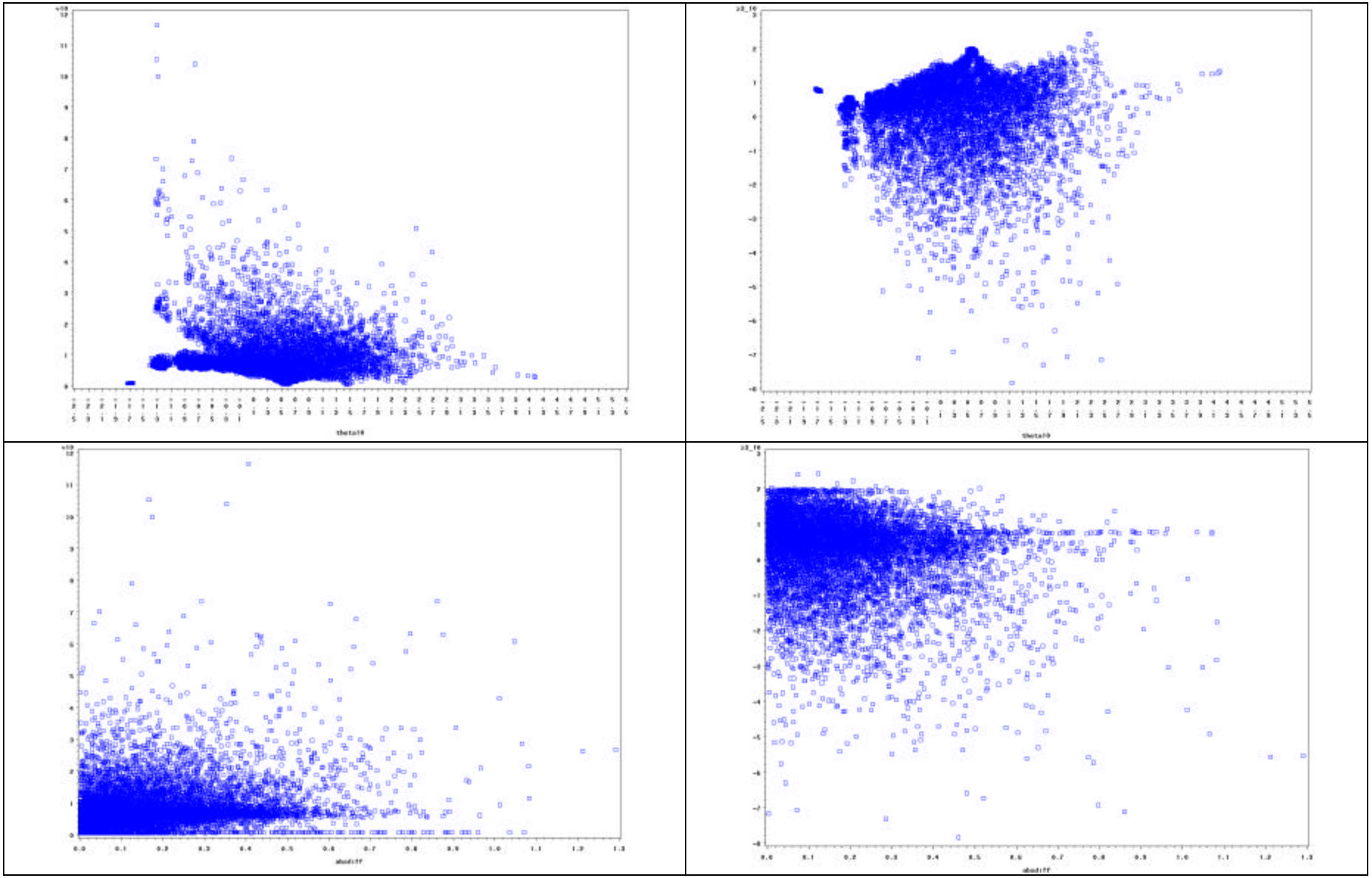

Figure 7. Scatterplots of 10-item theta scores (top, x-axis) and absolute-value differences between 10- versus 40-item thetas (bottom, x-axis) by person-fit indices (left = v, right = z_3).