Paper presented at the 2001 (April) Annual Conference of the
Society for Industrial and Organizational Psychology, San Diego.

Bimodal Score Distributions and the MBTI: Fact or Artifact?

Tammy L. Bess and Robert J. Harvey
Virginia Polytechnic Institute and State University

We examined Myers-Briggs Type Indicator (MBTI) score distributions computed using item-response theory (IRT) to assess the generalizability of earlier bimodality reports which have been cited in support of the "type" versus "trait" view of personality. Using the BILOG IRT program to score a sample of approximately 12,000 individuals who participated in leadership development programs, θ score distributions for the four dimensions of the MBTI computed using 10 (the BILOG default) versus 50 quadrature points were compared. Results indicated that past reports of bimodality were artifacts caused by BILOG's default use of a small number of quadrature points; when larger numbers of points were used, score distributions became strongly center-weighted. Although our findings are not supportive of the "type"-based hypothesis, the extremely high correlations between θ scores (rs > .996) suggest that no practical differences would be expected as a function of the number-of-quadrature-points decision.

The Myers-Briggs Type Indicator (MBTI) remains highly popular in applied organizational settings (e.g., Myers, McCaulley, Quenk, & Hammer, 1998), despite the fact that its Jungian "type"-based view of personality is conceptually quite different from the "trait" or continuous dimensional view of personality popularized by other instruments, and in the Five Factor Model (FFM) view of the structure of personality (e.g., Goldberg, 1992; McCrae & Costa, 1987, 1989). Although the MBTI has always been scored to produce continuous "preference scores" for each of its four dimensions – and indeed, in its most recent revision the MBTI adopted a continuous dimensional scoring system based on the 3-parameter logistic item response theory (IRT) model – its developers continue to emphasize the necessity of using dichotomous types when making assessment decisions, and not the continuous scores on the dimensions themselves that form the basis for the dichotomized types.

The MBTI was founded on Jung's (1921/1979) theory of psychological type, and focuses on the four dichotomies that were implicit or explicit in Jung's theory. Jung originally proposed that differences in behavior can be attributed to whether people are introverted or extraverted (the E-I dimension in the MBTI, which parallels the Extraversion scale of the FFM); this distinction was based on whether a person's energies were primarily directed toward the inner world of thought and experience (introversion), or oriented towards other people and situations (extraversion). Later, Jung extended his theory to include two additional dichotomies: the sensation (or sensing)-intuition (S-N, which parallels the FFM's Openness) distinction focused on the mental functioning of perceiving, and the thinking-feeling dichotomy (T-F, analogous to the FFM's Agreeableness) focused on the judging functions. During the development of the MBTI (e.g., Myers & Briggs, 1962), the judging-perceiving (J-P, analogous to the FFM's Conscientiousness) dichotomy – which was largely implicit in Jung's theory – was added, based upon Myers' unpublished typological work (e.g., see Myers et al, 1998).

Despite the strong theoretical foundation and the fact that the MBTI has been widely used, debate is ongoing with respect to the question of whether discrete personality types actually exist (e.g., Block & Ozer, 1982; Mendelsohn, Weiss, & Feimer, 1982; Miller & Thayer, 1989; Stricker & Ross, 1964). Unfortunately, at least with respect to the traditional preference-score method of scoring the MBTI, research has consistently shown that the bimodal score distributions implied by the "type" view of personality are <u>not</u> typically present in large, unselected populations of examinees. Although the absence of bimodal score distributions does not necessarily prove that the "type"-based approach is incorrect, if such distributions were to be found, this fact would definitely be cited as support for the MBTI's underlying type-based approach. Indeed, Myers and McCaulley (1985, pg. 157) agreed on the "simple attractiveness" of being able to find bimodal distributions.

The advent of an IRT-based scoring procedure for the MBTI (Harvey & Murry, 1994) was significant for a variety of reasons, with one unintended effect being that instead of producing the strongly center-weighted score distributions that had long been seen for the preference-score method, the θ (theta) scores produced by IRT instead exhibited a definite bimodal nature, with a relatively low density of subjects scoring in the middle of the distribution (i.e., at the type dichotomization point). This fact has been cited in support of the dichotomous type-based feedback model used by the MBTI (e.g., Myers et al., 1998). However, although Harvey and Murry (1994) did report that the IRT scoring system produced bimodal distributions, they also noted that the distributions were not especially sharply bimodal (p. 126), and stopped short of drawing the conclusion that their findings provided definitive evidence regarding the "type" versus "trait" controversy, calling instead for additional research in new samples.

Our study responded to theis need for a further investigation of the bimodality issue with respect to IRT scoring of the MBTI. First, although the Harvey and Murry

(1994) sample was sizable, it was not especially diverse, with college students representing the majority of the approximately 1,600 participants. Although college students clearly are people, and they arguably represent acceptable subjects for personality research, one can nevertheless question the degree to which results – especially, unprecedented results – obtained in a student-dominated sample would generalize to other populations (in particular, samples composed of individuals who complete the MBTI in a work-related, rather than a research-based context).

Second, it is possible that the bimodality reported by Harvey and Murry (1994) might have been an unintended artifact of the specific IRT scoring methodology used in that study. That is, Harvey and Murry (1994) used the default scoring parameters for EAP scoring with a normal prior distribution provided by the BILOG program (Mislevy & Bock, 1990) when estimating the θ scores. Although the BILOG program defaults would presumably be reasonable, one parameter in particular – the number of quadrature points used during the scoring process – may have had the potential to influence the shape of the distributions of θ scores. That is, in the context of binary IRT models, item-response probabilities are expressed as a joint function of the IRT scoring parameters *a, b,* and *c* for each item, at a given value of θ (*D* is a scaling constant, typically set to 1.702):

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad [1]$$

To estimate a θ score for each individual, a maximum likelihood function can be derived from the expected item-endorsement probability specified in [1] as follows (*n* = number of items, *u* = item response, and *Q* = 1-*P*):

$$l\ (\mathrm{u} \mid \theta) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i}$$
$$\ln l\ (\mathrm{u} \mid \theta) = \sum_{i=1}^{n} \left[(u_i \ln P_i) + \left[(1 - u_i) \right] \ln (1 - P_i) \right] \qquad [2]$$

(In practice, the log of the likelihood value is usually maximized instead of directly attempting to maximize the raw likelihood); here, *u* denotes the 1/0 response for each item (1 if the response was in the I, N, F, or P direction, and 0 otherwise). In BILOG, the θ score that maximizes the likelihood (or log-likelihood) function is computed using a quadrature-point method:

$$\theta = \frac{\sum_{k=1}^{q} x_k L(x_k) w(x_k)}{\sum_{k=1}^{q} L(x_k) w(x_k)}$$
$$PSD(\theta) = \left[\frac{\sum_{k=1}^{q} (x_k - \theta)^2 L(x_k) w(x_k)}{\sum_{k=1}^{q} L(x_k) w(x_k)} \right]^{1/2} \qquad [3]$$

In this method, the likelihood function is evaluated at only a small subset of *q* equally spaced points *x* along the θ scale (i.e., the quadrature points), each with an associated quadrature weight (*w*), and the θ estimate is computed simply by weighting and summing these values across the quadrature points (the PSD indexes the standard error of the θ estimate).

Although the choice of the number of quadrature points might not necessarily be expected to produce appreciably different θ estimates, upon comparing the locations of the default quadrature points used by BILOG for the MBTI (i.e., at θ = -4.0, -3.1, -2.2, -1.3, -0.4, 0.4, 1.3, 2.2., 3.1, and 4.0) against the bimodal θ distributions reported by Harvey and Murry (1994), we were struck by the fact that the areas of highest density corresponded closely to the locations of the two middle quadrature points. Hence, the second objective of our study was to compare the earlier scoring method using 10 quadrature points against the θ scores produced using a much larger number of quadrature points (i.e., 50). By using a larger number of points, this second method would more closely approximate the results that would be obtained in a brute-force approach of evaluating the likelihood function, and allow for a different shaped distribution to emerge if it were empirically so disposed (i.e., by locating a number of points throughout the middle range of the θ scale, rather than having only two points spaced nearly a standard deviation apart, with no point at the middle of the θ scale).

Method

Participants

The MBTI responses used in the present study were obtained from a large, non-profit leadership training and development organization. After eliminating missing data (for each scale, profiles were discarded if they contained any missing responses for the items in that scale), the final datasets were as follows: the EI scale contained 11,789 subjects, TF scale contained 12,338 subjects, SN scale contained 12,195 subjects, and JP scale contained 12,316 subjects. All subjects were managers in organizations taking the MBTI as part of a leadership development program.

Procedure

All data were analyzed using BILOG version 3.07 (the version used in Harvey & Murry, 1994) as well as the newer BILOG-MG version 1.1c. In order to examine the effect on θ score distributions due to the number of quadrature points, scoring analyses were run using the default number of points for scales of the length seen in the MBTI (10), as well as using the maximum-possible number of points (50). After calibrating the items and scoring them using EAP and the normal prior distribution (a uniform prior distribution, i.e., plain unweighted maximum-likelihood estimation, was also evaluated, and highly similar results were produced; to conserve space, only the EAP results are reported), θ score frequency distributions were examined for each of the four

scales, and correlations were computed between the θ estimates for each scale.

Results

With regard to the question of whether the number-of-quadrature-points decision may have influenced the Harvey and Murry (1994) report of bimodal score distributions for the MBTI, the results in Figures 1, 3, 5, and 7 depict the univariate frequency distributions produced using the BILOG default of 10 quadrature points, whereas those in Figures 2, 4, 6, and 8 depict the distributions produced using the maximum of 50 quadrature points. As in the Harvey and Murry study, the θ distributions for the default quadrature points produced a clearly bimodal shape. Additionally, as in the earlier study, the locations of the primary modes are located quite close to the θ = –0.44 and 0.44 quadrature points, and less-pronounced secondary modes can be observed at approximately θ = –1.3 and 1.3 (i.e., the locations of the adjacent quadrature points).

However, comparisons of each scale's distribution computed using a large number of quadrature points (Figures 2, 4, 6, and 8) against those produced using the default make it quite clear that the decision regarding the number and location of the quadrature points exerts a very strong effect on the subsequent shapes of the MBTI θ score distributions. As was found in earlier research using the preference-score based MBTI scales (e.g., Harvey & Murry, 1994; Stricker & Ross, 1964), heavily center-weighted, non-bimodal distributions result when a higher degree of granularity is used when evaluating the likelihood function.

With regard to the practical significance of the quadrature points choice, correlations between the 10- versus 50-quadrature point θ estimates for the E-I, S-N, T-F, and J-P scales were r = .99808, .99793, .99635, and .99794, respectively, indicating that although the quadrature point decision clearly exerts a powerful effect on distribution *shape*, the bottom-line impact of it on the θ scores themselves is primarily one of selectively stretching or compressing the θ metric, without producing any meaningful change in the *ordering* of individuals on the θ scale. That is, as Figure 9 illustrates for the E-I scale, the θ scores produced using a higher number of quadrature points are essentially a nonlinear transformation of the θ scores produced using the smaller number of points. Similar results were observed for the remaining scales.

Discussion

Our findings indicate that the enthusiasm seen among advocates of the MBTI based on the bimodal score distributions reported by Harvey and Murry (1994) needs to be significantly tempered in light of the fact that across all four dimensions, the results from the present study indicate that the earlier reports of bimodality were essentially artifacts caused by the particular number (and location) of quadrature points used by default in BILOG. Although we do not conclude that the absence of bimodality necessarily *proves* that the MBTI developers' theory-based assumption of categorical "types" of personality is invalid, the absence of empirical bimodality in IRT-based MBTI scores does indeed remove a potentially powerful line of evidence that was previously available to "type" advocates to cite in defense of their position.

Fortunately, because the main effect of quadrature-point choice appears to be a relatively modest, and selective, shrinking-stretching of the θ scale around the location of each point, an overwhelmingly strong correspondence exists between the θ scores estimated using different numbers of quadrature points. Thus, as a practical matter, we are at a loss to envision a situation in which it would make much of a practical difference which method were used, given that the θ score estimates correlate in excess of r = .996 across all four MBTI scales.

References

Block, J., & Ozer, D. J. (1982). Two type of psychologists: Remarks on the Mendelsohn, Weiss, and Feimer contribution. *Journal of Personality and Social Psychology*, *42*, 1171-1181.

Briggs, K. C. & Myers, I. B. (1976). *Myers-Briggs Type Indicator: For F.* Palto Alto, CA: Consulting Psychologists Press.

Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment, 4*, 26-42

Harvey, R. J., & Murry, W. D. (1994). Scoring the Myers-Briggs type indicator: Empirical comparison of preference score versus latent-trait methods. *Journal of Personality Assessment, 62,* 116-129.

Mendelsohn, G. A., Weiss, D. S., & Feimer, N. R. (1982). Conceptual and empirical analysis of the typological implications of patterns of socialization and femininity. *Journal of Personality and Social Psychology, 42*, 1157-1170.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

McCrae, R. R., & Costa, P. T. (1989). Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *Journal of Personality, 57*, 17-40.

Miller, M. L., & Thayer, J. R. (1989). On the existence of discrete classes in personality: Is self-monitoring the correct joint to carve? *Journal of Personality and Social Psychology*, *57*, 143-155.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs type indicato*r. Palto Alto, CA: Consulting Psychologists Press.

Myers, I. B., McCaulley, M. H., Quenk, N. L., & Hammer, A. L. (1998). *Manual: A guide to the development and use of the Myers-Briggs type indicator*. Palo Alto, CA: Consulting Psychologists Press.

Sipps, G. J., Alexander, R. A., & Friedt, L. (1985). Item Analysis of the Myers-Briggs type indicator. *Educational and Psychological Measurement, 45*, 789-796.

Stricker, L. J., & Ross, J. (1964). Some correlates of a Jungian personality inventory. *Psychological Reports, 14*, 623-643.

Tzeng, O. C. S., Ware, R., & Bharadwaj, N. (1991). Comparison between continuous bipolar and unipolar ratings of the Myers-Briggs type indicator. *Educational and Psychological Measurement, 51*, 681-690.

Tzeng, O. C. S., Ware, R., & Chen, J. M. (1989). Measurement and utility of continuous unipolar ratings for the Myers-Briggs type indicator. *Journal of Personality Assessment, 53*, 727-738.

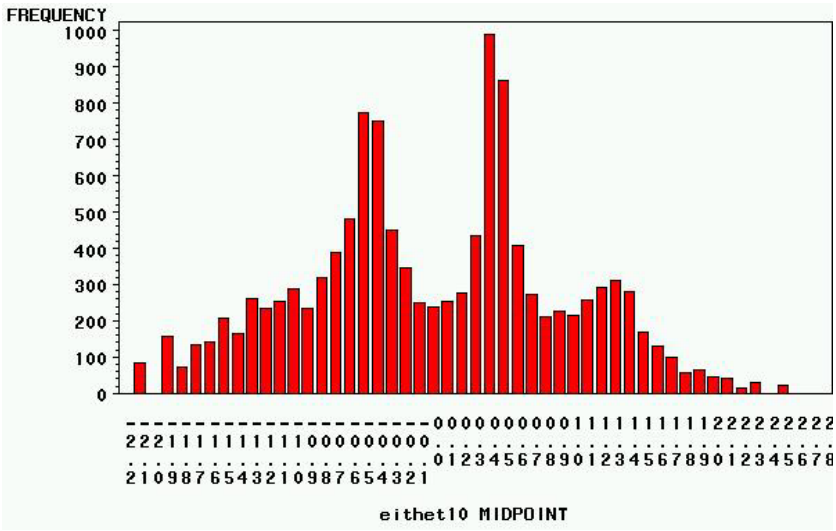

Figure 1 Frequency distribution of IRT theta score estimates using default quadrature points for the EI scale.

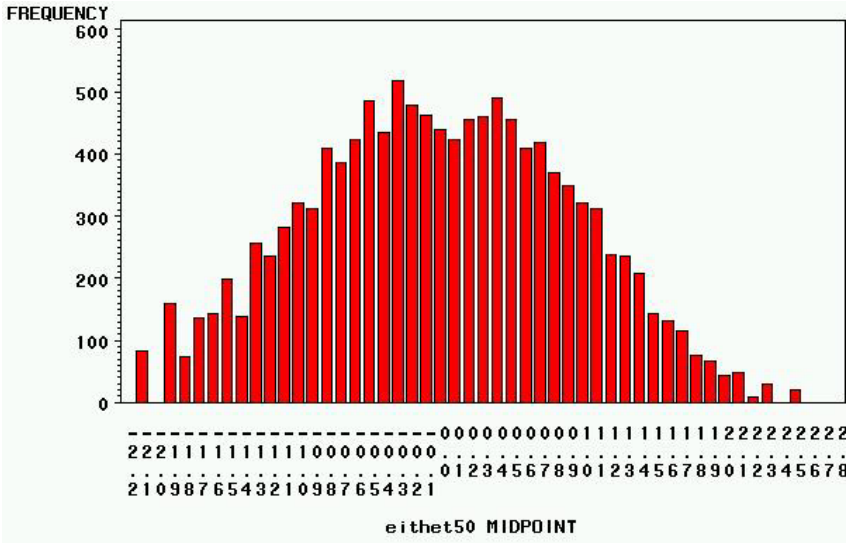

Figure 2. Frequency distribution of IRT theta-score estimates using fifty quadrature points for the EI scale.

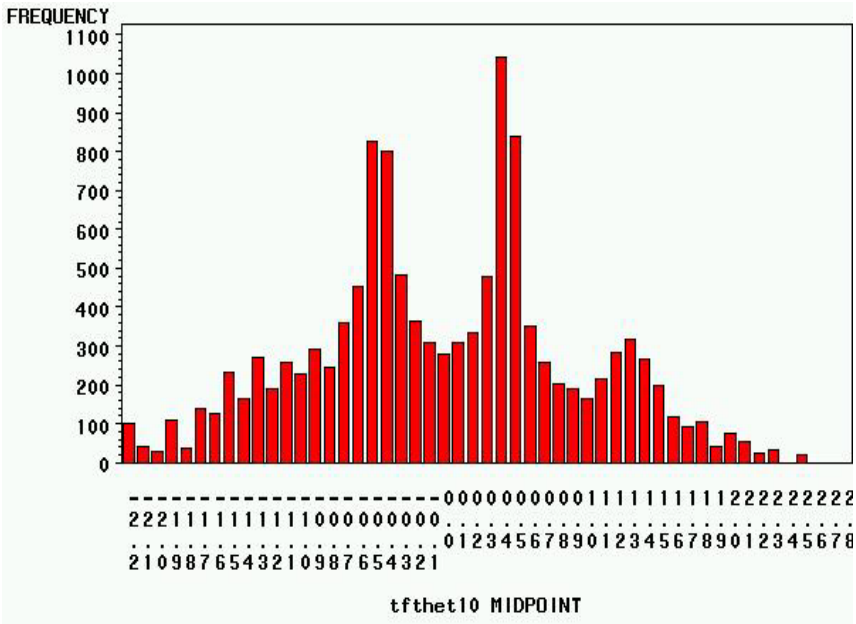

Figure 3. Frequency distribution of IRT theta score estimates using default quadrature points for the TF scale.

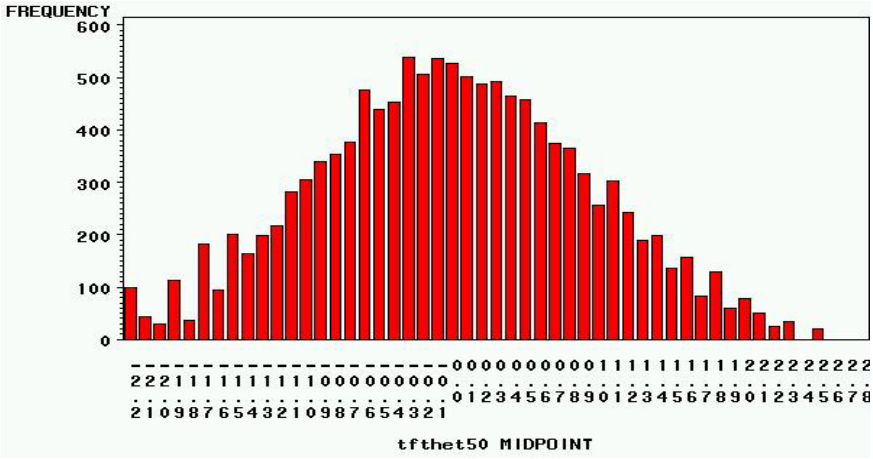

Figure 4. Frequency distribution of IRT theta-score estimates using fifty quadrature points for the TF scale.

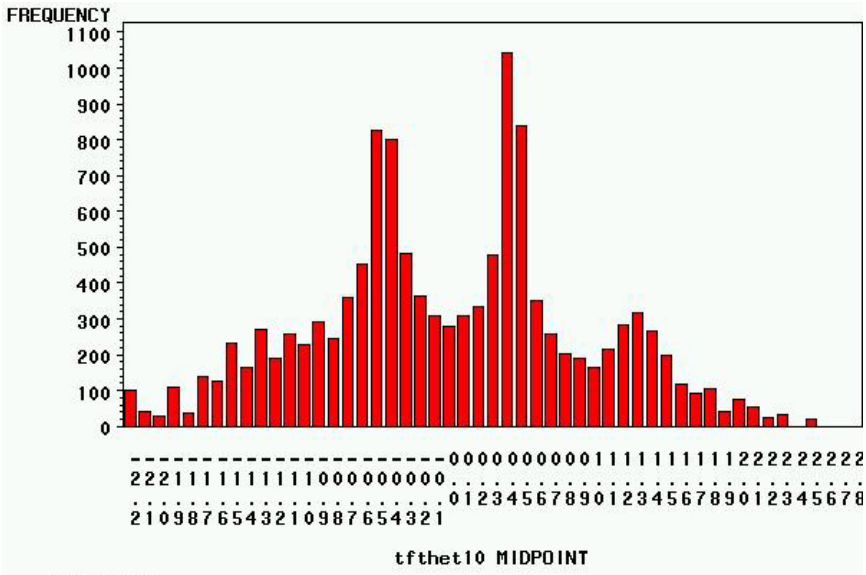

Figure 5. Frequency distribution of IRT theta score estimates using default quadrature points for the SN scale.

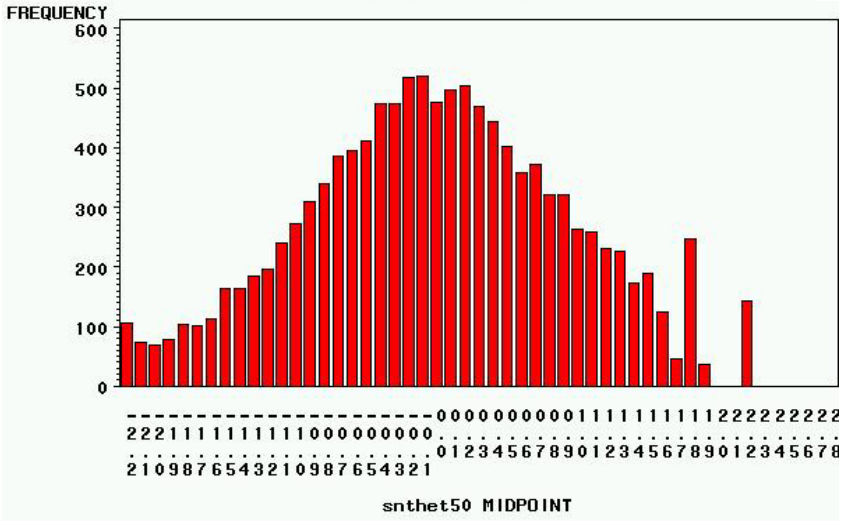

Figure 6. Frequency distribution of IRT theta-score estimates using fifty quadrature points for the SN scale.

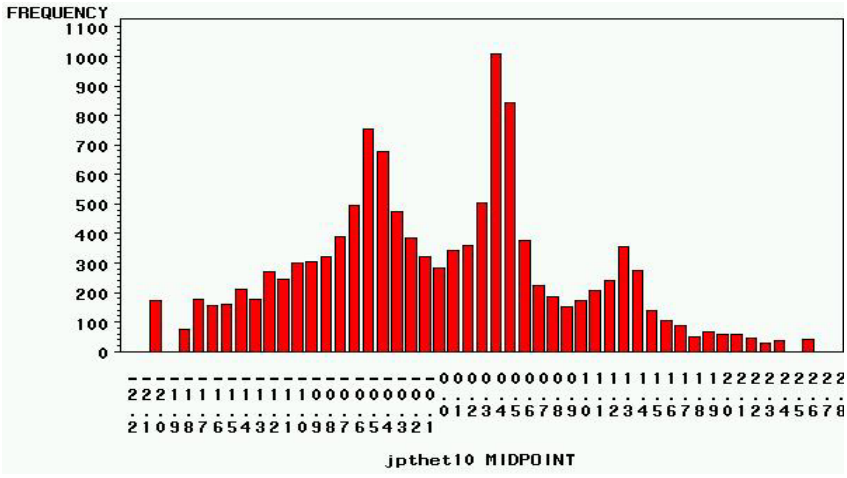

Figure 7. Frequency distribution of IRT theta score estimates using default quadrature points for the JP scale.

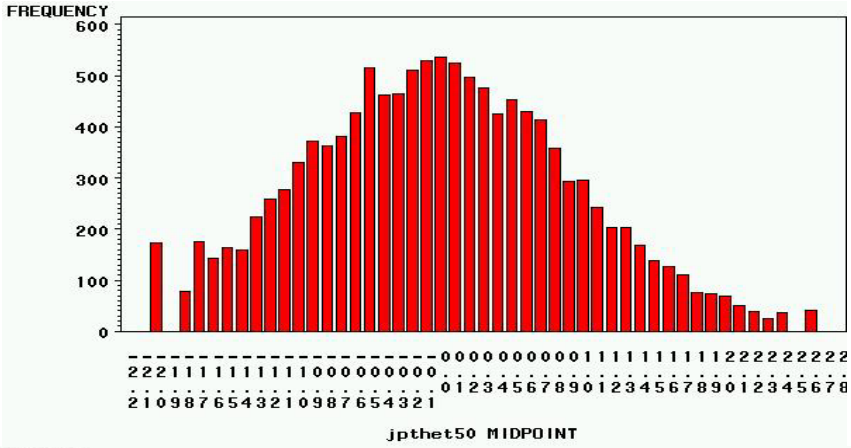

Figure 8. Frequency distribution of IRT theta-score estimates using fifty quadrature points for the JP scale.

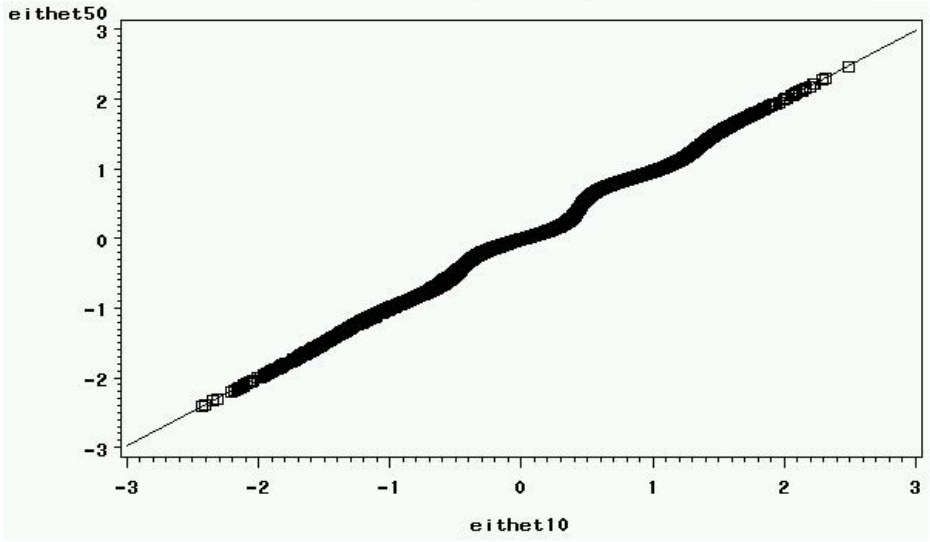

Figure 9. Scatterplot of the EI scale results computed using 10 versus 50 quadrature points.