DOCUMENT RESUME

ED 358 113 TM 019 879

AUTHOR	Wise, Lauress
TITLE	Test Form Accuracy.
INSTITUTION	Defense Manpower Data Center, Monterey, CA.
PUB DATE	Apr 93
NOTE	59p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Classification; *Item Response Theory; Military Personnel; Occupational Tests; Personnel Evaluation; *Profiles; *Scores; Test Construction; *Test Format; Test Use; Test Validity
IDENTIFIERS	*Accuracy; Armed Services Vocational Aptitude Battery; Calibration; *High Stakes Tests; Profile of American Youth; Target Planning

ABSTRACT

As high-stakes use of tests increases, it becomes vital that test developers and test users communicate clearly about the accuracy and limitations of the scores generated by a test after it is assembled and used. A procedure is described for portraying the accuracy of test scores. It can be used in setting accuracy targets durıng form construction and in communicating information to test users. The procedure is discussed in the context of the Armed Services Vocational Aptitude Battery. The general procedure for reviewing current test accuracy profiles and for setting new targets includes: (1) developing accuracy profiles for current fcrms; (2) using expert judgmenl to review and revise accuracy goals for new forms at key points on the target scale; (3) usir_ item response theory analyses to calibrate new items, adjust for sample differences, and estimate classification error rates for trial forms; (4) developing preliminary tolerances for compliance with accuracy targets; and (5) checking initial form accuracy profiles. The procedures for producing accuracy profiles are illustrated with 4,000 cases from the Profile of American Youth Study. Advantages and limitations of the procedure are discussed. Twelve two-part graphs illustrate the analyses. (SLD)

Test Form Accuracy

Dr. Lauress Wise
Defense Manpower Data Center

Paper presented at the Annual Convention of the National Council for Measurement in Education, April 13, 1993, Atlanta, GA.

NOTE: The views expressed are those of the author and do not necessarily reflect the position of the Defense Manpower Data Center or other agencies within the Department of Defense.

Test Form Accuracy

Dr. Lauress Wise
Defense Manpower Data Center

Paper presented at the Annual Convention of the National Council for Measurement in Education, April 13, 1993, Atlanta, GA.

NOTE: The views expressed are those of the author and do not necessarily reflect the position of the Defense Manpower Data Center or other agencies within the Department of Defense.

TA263288.CDR

DEFENSE MANPOWER DATA CENTER
99 Pacific Street, Suite 155A • Monterey, CA 93940

TEST FORM ACCURACY

How long should my test be? How hard should my test be? These are two very general questions that must be answered in building any test. The correct answers depend, of course, on how the information from the test is to be used.

Underlying the general questions of test length and difficulty is the general issue of how accurate the scores generated for each subtest in the form should be. This is a policy question, involving tradeoffs between the variable costs of test development and administration related to precision and the benefits from more accurate estimates of the underlying abilities.

High-stakes use of test scores is increasing, and so it is vital that test developers and test users communicate clearly about the accuracy and limitations of the scores generated by a test after it is assembled and used operationally. This paper describes a procedure for portraying the accuracy of test scores. The procedure can be used both in setting accuracy targets during form construction and in communicating information about score accuracy to test users after forms are put into operation.

Background

As part of a general review of the Armed Services Vocational Aptitude Battery (ASVAB), the item difficulty targets used in constructing new forms of the ASVAB are being examined. Two questions are under consideration. First, what should the target distribution of items difficulties be for each of the tests in the battery? Second, how should new forms be constructed to ensure adequate adherence to these targets?

In the past, the primary strategy for building essentially equivalent forms has been to match item difficulties to the reference form on an item-by-item basis. This procedure has generally been sufficient to produce new forms close enough in overall difficulty to the reference form so as to allow reasonable score equivalence through equating. The procedure places severe limitations on item development, however, as many good items cannot be used simply because they do not happen to match the difficulty of a reference form item. Item-by-item matching is not a necessary procedure. It is possible to construct forms of equivalent difficulty by matching at the level of the overall difficulty distributions.

Classical test theory (CTT) (e.g., Lord & Novick, 1968) describes test accuracy in terms of the reliability coefficient, an estimate of the correlation of scores from two parallel forms. This approach assumes that error of measurement is constant throughout the measurement scale. This assumption may not seem tenable, as it would seem that a test form with mostly easy items would be more accurate at the low end of the ability scale than at the high end. One must remember, however, that classical test theory was designed for use with a number right score. At every point in the scale, one unit corresponds to one more item correct so that the homogeneity of error assumptions may not be as reasonable as they appear. Nonetheless, several efforts have been made to estimate errors of measurement for specific number correct score levels (Qualls-Payne, 1992; Feldt, Steffen, & Gupta, 1985). More recently Kolen, Hanson, and Brennan (1992) have demonstrated an approach to computing conditional errors of measurement for transformed (scale) scores as well as for raw, number correct scores.

In general, a number right score may not be the best metric for consideration of difficulty targets. The relationship between an examinee's true ability and his or her number right score depends very heavily on the difficulty of each of the items in the test form. Two forms with different item difficulty distributions will have different number correct score distributions for any given sample or population. In the ASVAB program, standardized subtest scores are used as the basis for forming composites; for our most important composite, the Armed Forces Qualifying Test (AFQT) composite, we use percentile scores in making selection decisions.

Item response theory (IRT) has been advanced as an alternative to CTT, in part to counter scale-constancy issues that arise with use of a number right scale. Latent ability is scaled so that the regression of each item score on the underlying ability follows a fixed functional form, usually a normal ogive or three-parameter logistic (3PL) function. Using IRT models, it is possible to estimate the accuracy of the score obtained by each individual at a given underlying ability level as a function of characteristics of the items used in the measurement. Thus, accuracy is viewed to vary across the measurement scale and can be estimated from item parameters (Lord 1980; Lord & Novick, 1968). Lord (1984) provided an approach that uses IRT concepts to estimate score accuracy when scores are based on the number of correct responses rather than a direct (maximum likelihood or Baysian) estimate of underlying ability.

Score and Accuracy Metrics

In describing test form accuracy and setting accuracy standards, we determine both a metric for describing score levels and a metric for portraying accuracy at each of these levels. With respect to score-level metrics, unfortunately, neither the IRT theta metric nor the number correct metric is used with the ASVAB in making personnel decisions. A commonly used metric is a percentile scale where each examinees score is compared to the distribution of scores from a fixed reference population. For the ASVAB, the 1980 Youth Population (OASD, 1982) is used as the primary point of comparison both for the high school career exploration program and for making operational enlistment eligibility decisions.

The youth population percentile metric has been selected for portraying score-levels for two reasons. First, this is the metric used in the general determination of qualification. A second reason is that accuracy judgments should be linked to some population distribution. We should be relatively unconcerned about accuracy at points in the scale where there are few individuals to be evaluated and much more concerned at those points where many examinees will score. For the percentile metric, the relative number of examinees at scoring at each point is essentially the same. (About two percent of the relevant population will score within one point of any given level.)

Given the choice of the percentile metric for describing examinee abilities, what metric should be used to describe accuracy? Typically, a standard error, defined as the expected standard deviation of an individual's scores across parallel forms (overall or at particular score levels), is used as a measure of accuracy. Alternatively, the distance between specified percentile points (confidence bound cutoffs) in the conditional distribution of observed scores given a "true" score might be used as the measure of accuracy.

We are currently pursuing a different metric for describing accuracy. The primary use of the test scores is to classify applicants, dichotomously, as either qualified or not qualified (overall or for a particular job). Consequently, we are using classification error rates as the measure of score accuracy. The classification error rate is defined as the proportion of examinees will be incorrectly classified, either as qualified when they are not (false positives) or as unqualified when they actually are (false negatives). A classification error rate metric communicates the operational impact of score accuracy, and may be more appropriate than standard error measures when communicating with policy makers. (Please note that in this context, true qualification is defined in terms of the trait being *measured* by the test and not by some more ultimate criterion that is being *predicted* by the test.)

Approach

As described above, we are using classification error rates for defining and communicating score accuracy. The general procedure for reviewing current test accuracy profiles and for setting new targets as appropriate includes: (1) develop accuracy profiles for current forms; (2) use expert judgment to review/revise accuracy goals for new forms at the key points (ranges) on the target scale; (3) use IRT analyses to calibrate new items, adjust for differences between the tryout sample and the Youth Population, and estimate classification error rates for trial forms during form assembly; (4) develop preliminary tolerances for compliance with accuracy targets;

and (5) check the initial form accuracy profiles against revised accuracy profiles computed on operational samples during formal equating, and revise targets/tolerances as required.

The remainder of this paper presents details of the procedures for obtaining item parameter estimates and using them to generate accuracy profiles for actual and potential forms. These procedures are illustrated with analyses of data from the Profile of American Youth Study (OASD, 1982).

Samples

The Profile of American Youth Study (OASD, 1982) provided the basis for the current ASVAB norms. It involved administration of the ASVAB reference form to a complex sample of approximately 12,000 youth. We drew a systematic sample of 4,0C0 cases from the data files using sampling probabilities that were inversely proportional to their current sampling weight. The overall selection probability was thus the original selection probability (the inverse of the sampling weight) times the probability of being selected for this new subsample (the weight times a constant). Thus composite probability was a constant and the data could be analyzed without having to use case weights.

We next divided the 4,000-case sample into two 2,000-case samples (alternating in order of selection into the 4,000-case sample) for cross-validation purposes (and because we were using a PC version of BILOG to get item parameter estimates). The result of all of these machinations was two 2,000-case samples that were each representative of the entire youth population without having to use differential case weights.

For illustrative purposes, we examined the Word Knowledge (WK) and General Science (GS) tests. WK is notorious for having an abundance of relatively easy items, while GS is more balanced with respect to item difficulty.

Methods

IRT parameter estimates. We obtained item parameter estimates for each of the two tests in each of the two 2,000 case samples. The BILOG program was used with options specifying floating priors for the slope and asymptote (a and c) parameters and no prior for the threshold (b) parameters. The "Free" option was specified to allow discrete estimation of the marginal population distribution, and the number of quadrature points was increased to 31. We also estimated individual scores (theta) using EAP estimation with standard normal priors and rescaling (both item and subject parameters) so that the latent population distribution (rather than the observed sample distribution) would have mean 0 and standard deviation 1. In the "SCORE" step, 26 quadrature points were used.

If this were not a strictly representative sample from the Reference Population (RP), we would have to adjust the item parameter estimates for differences between the calibration sample and the RP. Typically, the reference form is administered to a sample that is randomly equivalent to the sample used to calibrate new items. Differences in reference form item parameter estimates from the youth population sample and the new sample provide the basis for translating the new item parameter estimates back onto the reference population theta scale. One approach, for example, is to find the linear transformation that minimizes the (weighted) average squared difference in the test characteristic curves based on the original calibration and the rescaled new estimates (Stocking & Lord, 1983). Alternatively, the differences to be minimized may be expressed relative to the estimated standard error of the differences (jointly for the slope and threshold parameters) defined in terms of a chi-square test statistic (Divgi, 1985).

Percentile to theta translation. In computing test accuracy for a particular score level, we need to know the theta value (underlying, true ability) corresponding to each score level in order to compute expected observed score distributions (using an IRT model) and then classification errors. We examined three ways of linking theta and percentile scores. These were: (1) assume a normal distribution on the latent (theta) scale and use the inverse of the cumulative normal distribution function to map percentiles onto theta; (2) compute the distribution of theta score estimates in the youth population samples and use the inverse of this empirical cumulative distribution function; and (3) sum the posterior theta densities for the youth population sample examinees and compute a cumulative distribution function based on this composite posterior theta density. Figures 1 and 2 compare the resultant percentile-to-theta functions from each of these three methods for each of the two samples and each of the two tests. Figures 3 and 4 show the differences in estimated thetas at each percentile level. Each of the three methods led to very similar results, except at the extremes. The "observed" theta distribution method (method 2) led to the most diverse results at the extremes. We continued with the results from method 3 which estimated the cumulative distribution of the underlying (true) theta values rather than the distribution of theta estimates.

Conditional expected observed score distributions. For each percentile point (from 0.5 to 99.5 in increments of 1), we identified the corresponding theta value and used our item parameter estimates and the 3PL IRT model to estimate a probability of passing for each item given that theta value. The conditional distribution of the number correct score on theta, under local independence, is a compound binomial distribution (see Lord, 1984) which is difficult to express in closed form, but not difficult to compute. The compound binomial distribution may be expressed recursively as follows. If $P_k(x)$ is the probability of x correct on the first k items (for each x from 0 to k), then the probability of x correct after k+1 items is $(1-p_{k+1})*P_k(0)$ if x=0 and otherwise is given by $P_{k+1}(x) = (1-p_{k+1})*P_k(x) + p_{k+1}*P_k(x-1)$, where p_{k+1} is the probability of a correct response on the item k+1 given theta. Figures 5 and 6 show the mean and the 5th and 95th percentile points of the conditional-number-correct distributions for each percentile level.

The next step was to convert the conditional distributions from a number-correct metric to a percentile-score metric. We computed the cumulative distribution of number-correct scores for each test in each of the two reference population samples. Figures 7 and 8 show these distributions. Next, we used these cumulative distribution functions to convert each number-correct score to a percentile score and applied these conversions to the number-correct scores in the conditional expected-score distributions. We thus had estimates of the probability of obtaining each possible *estimated* percentile score for a given *true* percentile score. Figures 9 and 10 show the mean and 5th and 95th percentile for the conditional expected score distributions after converting these distributions to an estimated youth population percentile metric.

Compute classification error rates. Numerical Integration (using the 100 discrete percentile levels) was used to compute expected classification error rates. Target classification levels were defined by alternative cut scores on the underlying (true) percentile metric. We examined classification levels varying from the 1st to the 99th percentile. For each target classification level, we summed the probabilities of a conditional *estimated* (observed) score that was above the classification level across all *true* percentile levels that were *below* the target to estimate the proportion (rate) of false positives. Similarly, we computed the proportion of false negatives as the likelihood that an examinee will have a *true* percentile level below the target level, but have an *estimated* percentile above the target. We then summed the false positive and false negative proportions to get the total classification error rate for each target point on the percentile scale. (Actually, our computer did most of the summing.) The resulting accuracy profiles for Reference Form WK and GS tests are shown in Figures 11 and 12.

The "scallop patterns" shown in Figures 11 and 12 were not fully expected, but easy to explain due to the discrete nature of number correct to percentile conversions. With only 26 or 36 possible raw scores, it is not possible to obtain each of the 99 possible percentile scores. When a classification level (cut score) matches an obtained percentile score, there is maximal uncertainty about examinees who receive that percentile score. Half of the time they will actually be above the cut score and half the time they will be below it. On the other hand, when the cut score falls midway between two obtained percentile scores, uncertainty about examinees with scores near the cut point is minimized.

One feature of the classification error rate profiles that was fully anticipated was the tendency for error rates to be highest in the middle of the percentile range and to drop to zero at the endpoints. When a cut score is in the middle, a significant number of the examinees will have scores near the cut score leading to potential classification errors. When the cut scores are near the top (bottom) of the distribution, the majority of the examinees will have true scores that are far below (above) the cut score leading to minimal chances for errors.

Differences between the GS error profiles in Figure 11 and the WK error profiles in Figure 12 also were expected. WK is a longer test with 35 items, compared to 25 items for GS. This difference in length leads to higher levels of reliability and lower maximum error rates. WK has a maximum error rate of about 9 percent, compared to a maximum rate of 12 percent for GS. Also, the error rate profile for WK has a negative skew with relatively lower error rates for lower score levels. This is consistent with the fact that WK has a lot of easy items, so higher accuracy would be expected at the lower end of the ability distribution where the easy items provide the most information. By contrast, GS has a more even mix of item difficulties and also a more symmetric error rate profile. These results illustrate the power of the error profile approach to portray important consequences of different item difficulty mixes.

Summary

The results of these illustrative analyses demonstrate the feasibility of using expected classification error rates to assess the consequences of different mixes of item difficulty and discrimination levels. If this is so, we will not need to continue with item-by-item, "p- value" matching. Given initial development of percentile to theta conversions, it takes 10 to 15 minutes to go from a set of item parameter estimates to classification error plots (most c the time is importing and formatting the results in Harvard Graphics), so iterative use of this approach with alternative item sets appears feasible.

Some limitations of this approach should also be noted. The accuracy of the accuracy portrayal rests on the appropriateness of the two key item response theory assumptions: (1) the item characteristic curves (giving the probability of passing as a function of underlying ability) are reasonably approximated by the three-parameter logistic regression function and (2) the probabilities of passing different items are independent for a given ability level (local independence). Both of these assumptions are used in computing the conditional expected score distributions.

Preliminary discussions with policy makers have reinforced the value of the error profile approach. The error profiles communicate more concretely the impact of errors of measurement, making it easier to defend the additional testing time required to achieve greater accuracy levels. Whether error profiles will prove helpful in determining initial test specifications remains to be seen, but their value in determining the degree to which alternative forms provide equivalent measures is obvious.

References

Divgi, D.R. (1985). A minimum chi-square method for developing a common metric in item response thυory. *Applied Psychological Measurement, 9*, 413-416.

Feldt, L.S., Staffen, M., & Gupta, N.C. (1985). A comparison of five models for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351-361.

Kolen, M.J., Hanson, B.A., & Brennan, R.L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29,* 285-307.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum.

Lord, F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21,* 239-243.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

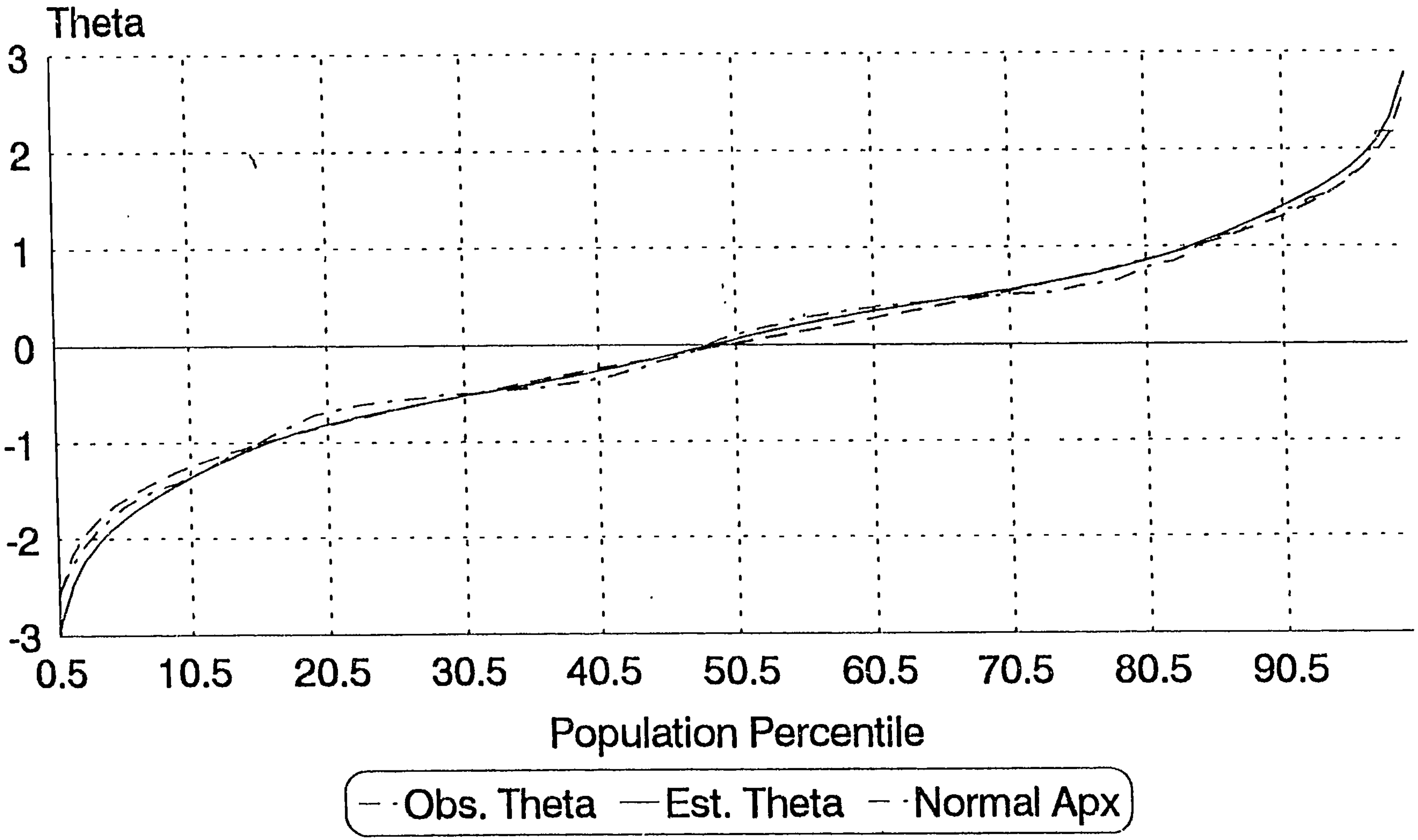
Office of the Assistant Secretary of Defense. (1982). *Profile of American youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery.* Washington, DC: Department of Defense.

Qualls-Payne, A.L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement, 29,* 213-225.

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Figure 1a. Estimated Theta by Population Percentile

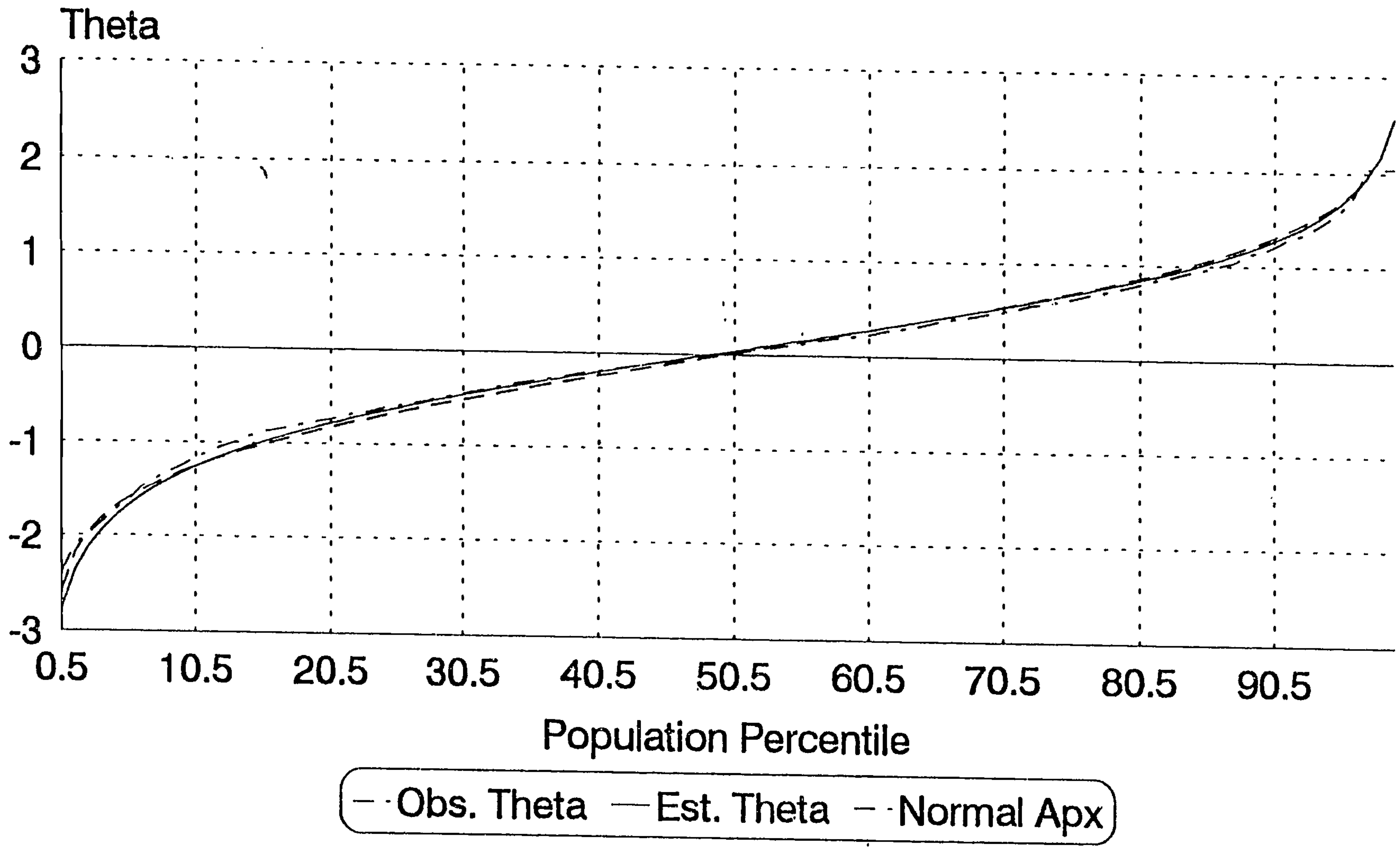
Reference Form GS Subtest



Based on First 2,000 Case, Self-Weighted Sample

Figure 1b. Estimated Theta by Population Percentile

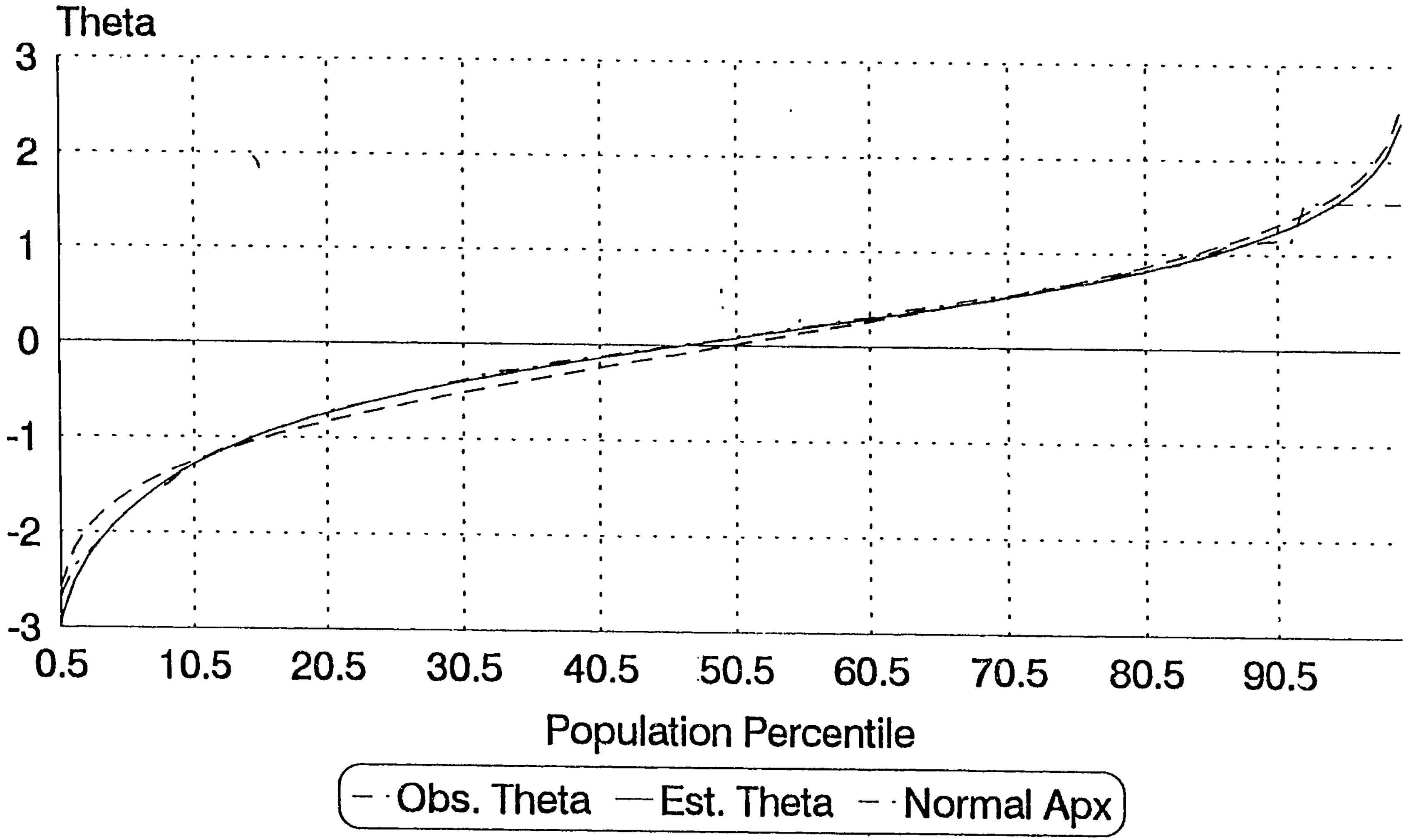
Reference Form GS Subtest



Based on Second 2,000 Case, Self-Weighted Sample

Figure 2a. Estimated Theta by Population Percentile

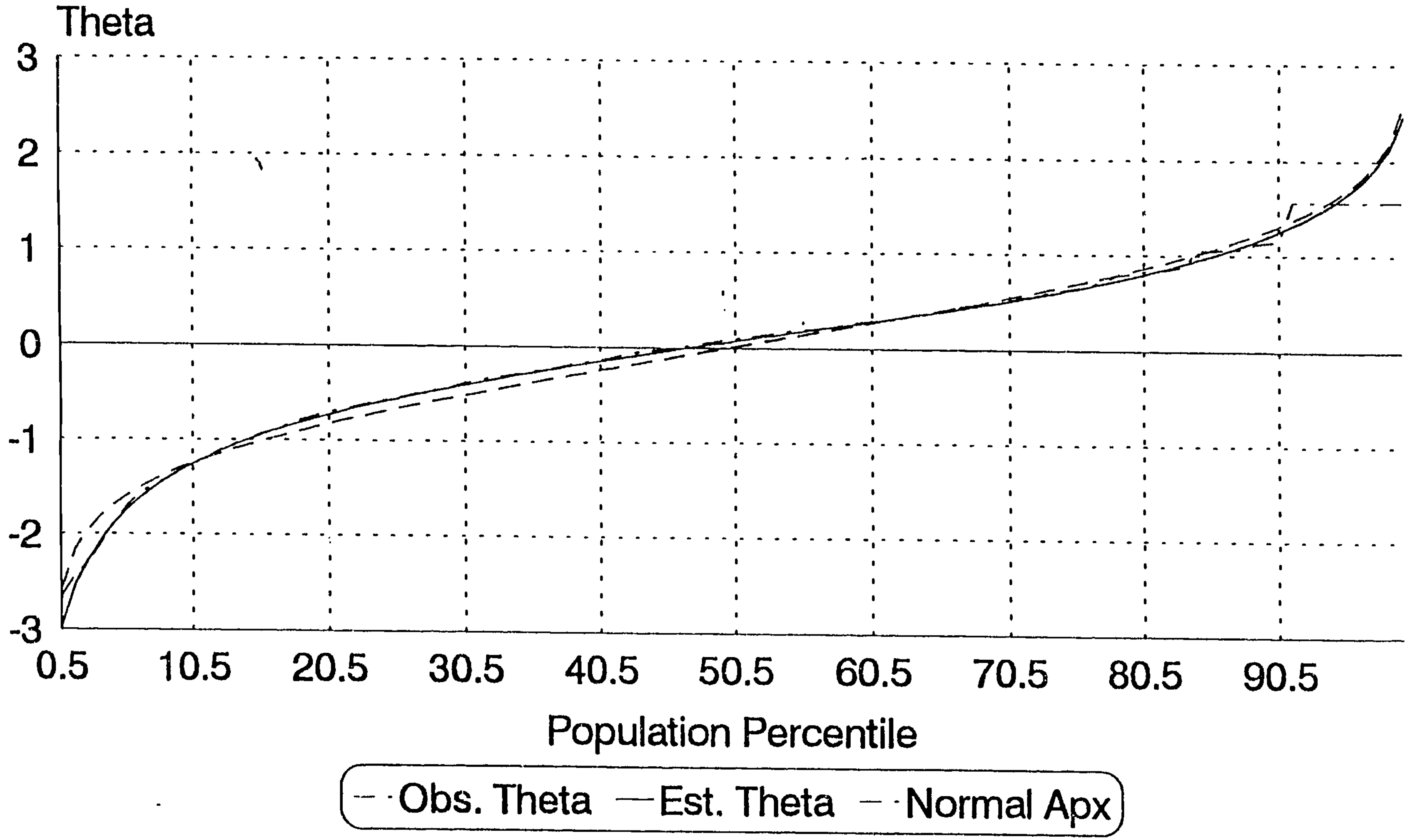
Reference Form WK Subtest



Based on First 2,000 Case, Self-Weighted Sample

Figure 2b. Estimated Theta by Population Percentile

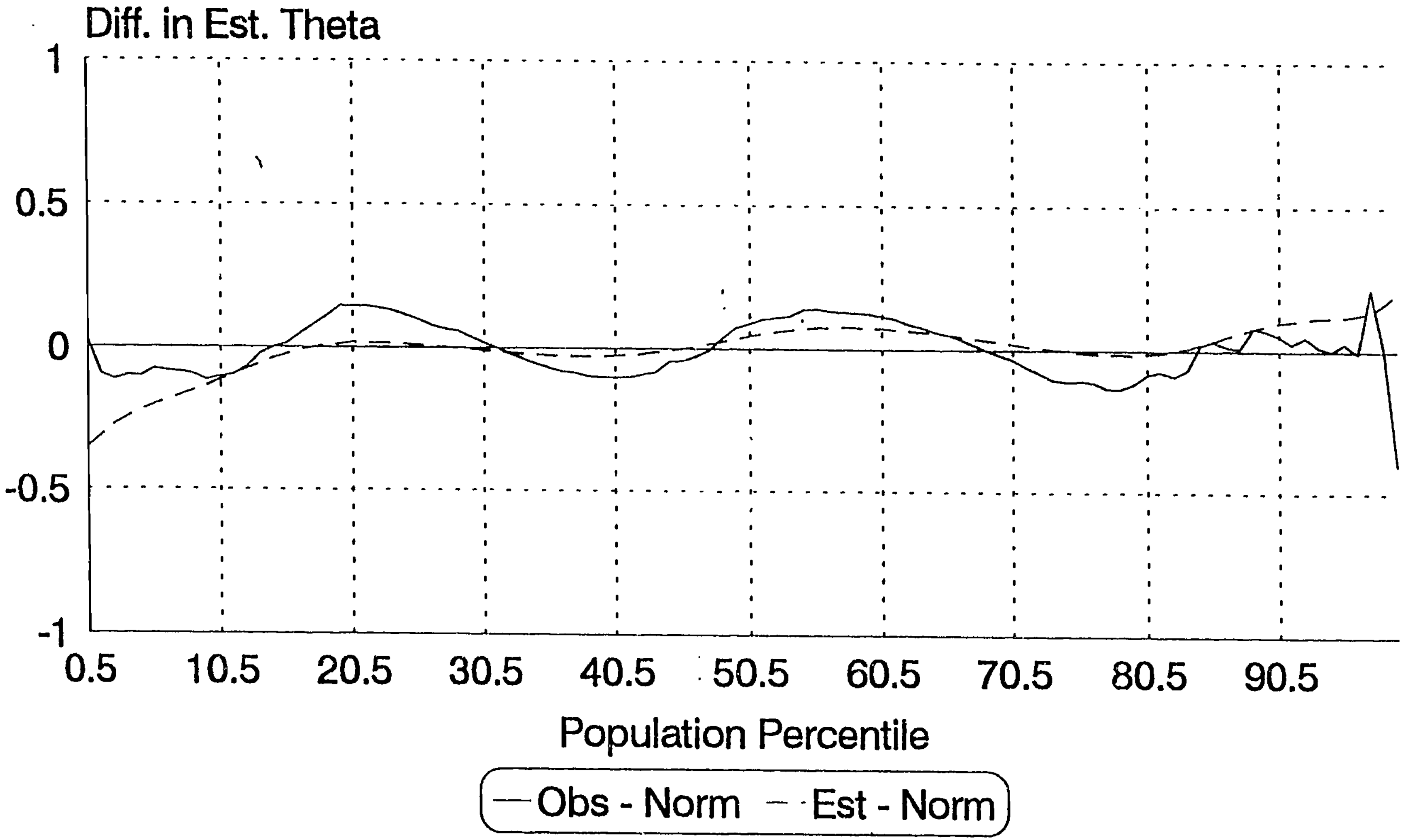
Reference Form WK Subtest



Based on Second 2,000 Case, Self-Weighted Sample

Figure 3a. Difference in Estimated Theta for Population Percentiles

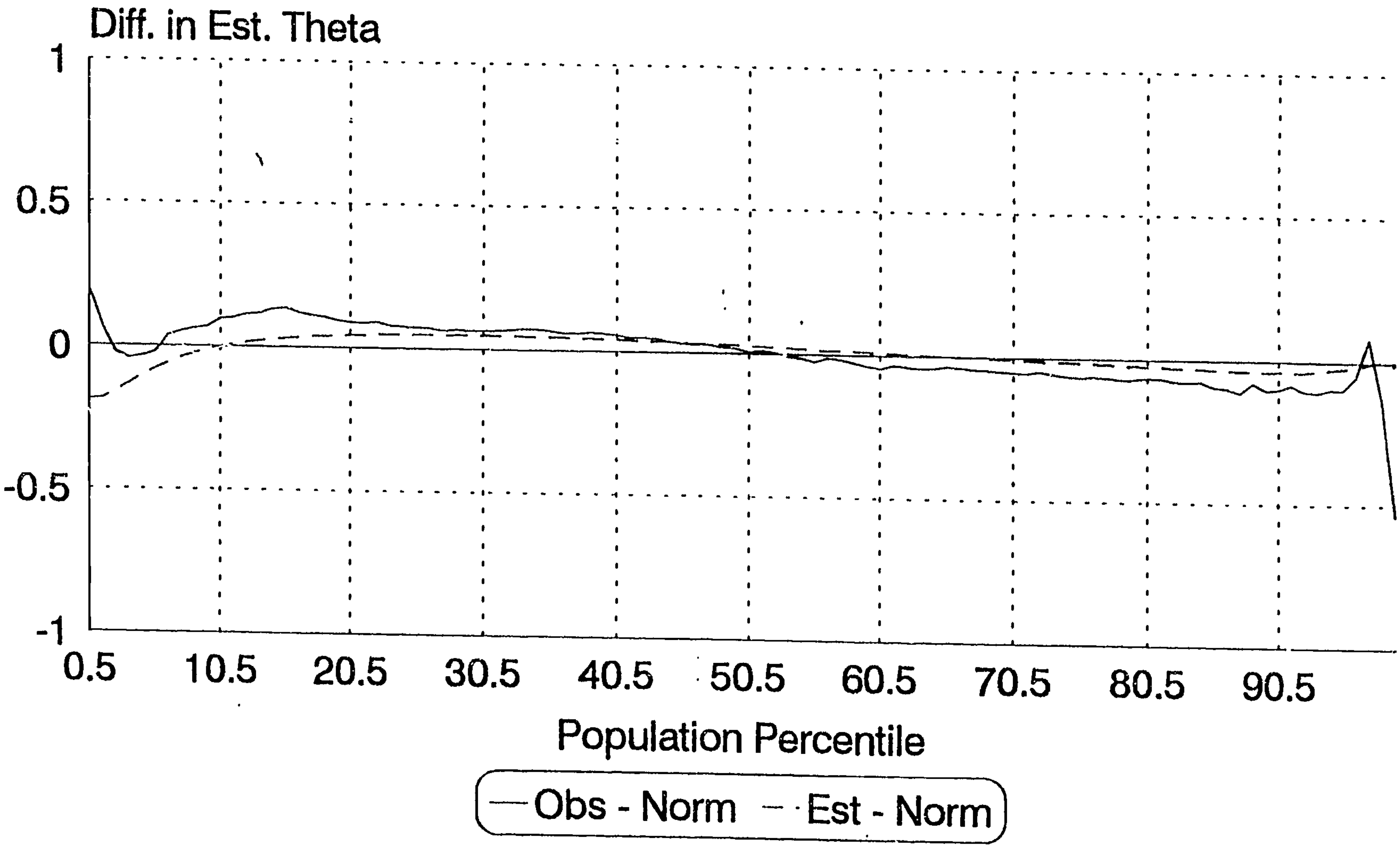
Reference Form GS Subtest



Based on First 2,000 Case, Self-Weighted Sample

Figure 3b. Difference in Estimated Theta for Fopulation Percentiles

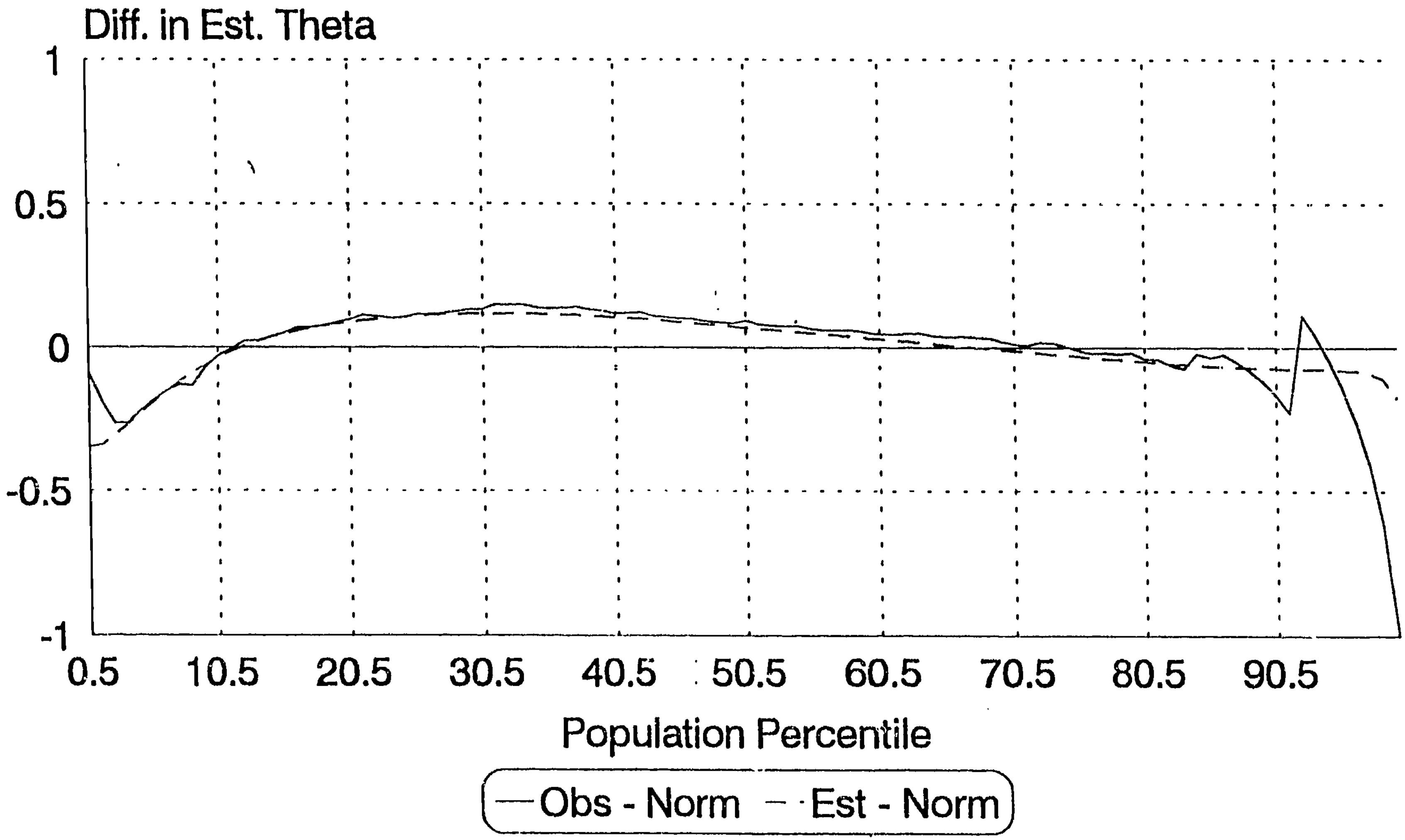
Reference Form GS Subtest



Based on Second 2,000 Case, Self-Weighted Sample

Figure 4a. Difference in Estimated Theta for Population Percentiles

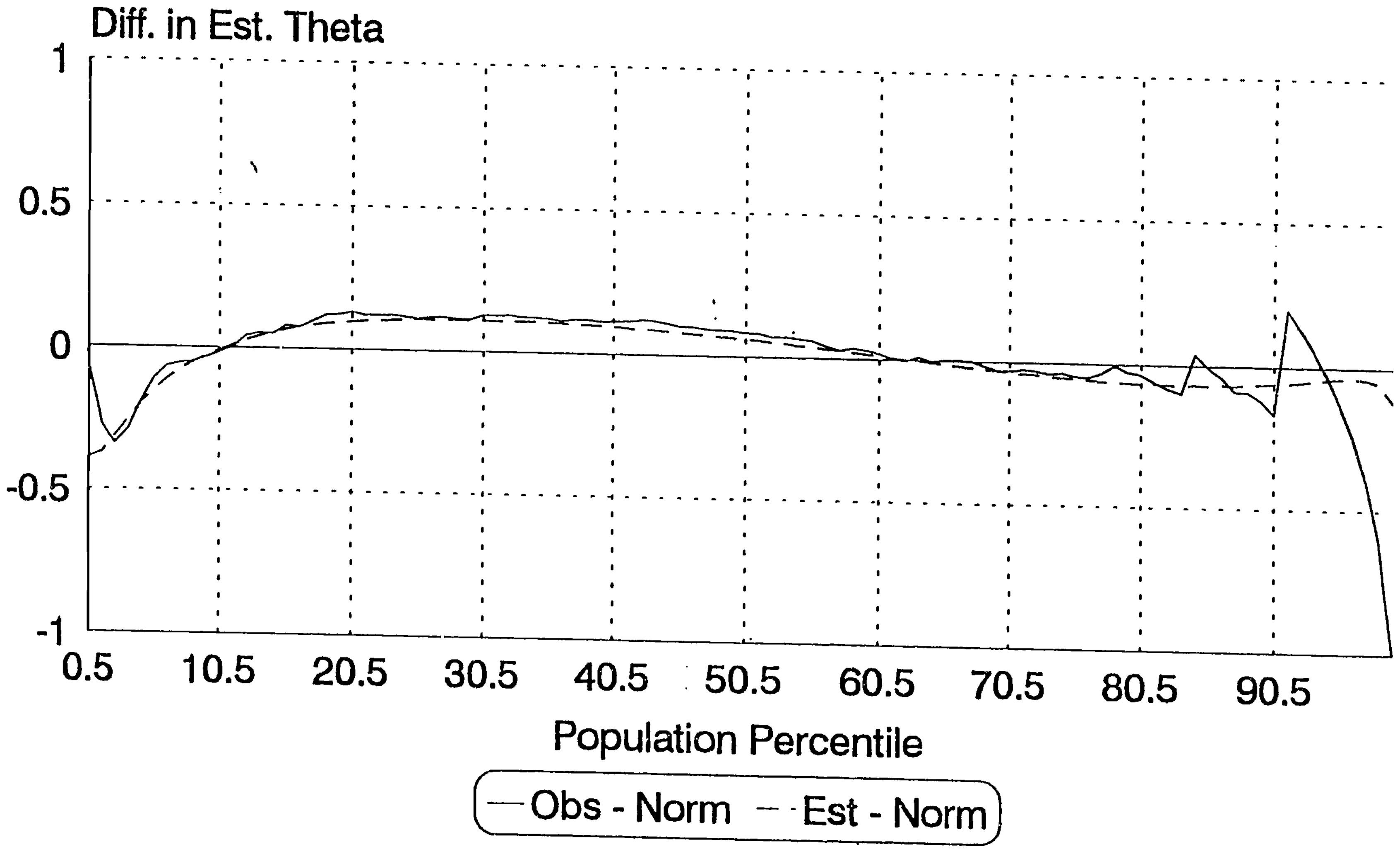
Reference Form WK Subtest



Based on First 2,000 Case, Self-Weighted Sample

Figure 4b. Difference in Estimated Theta for Population Percentiles

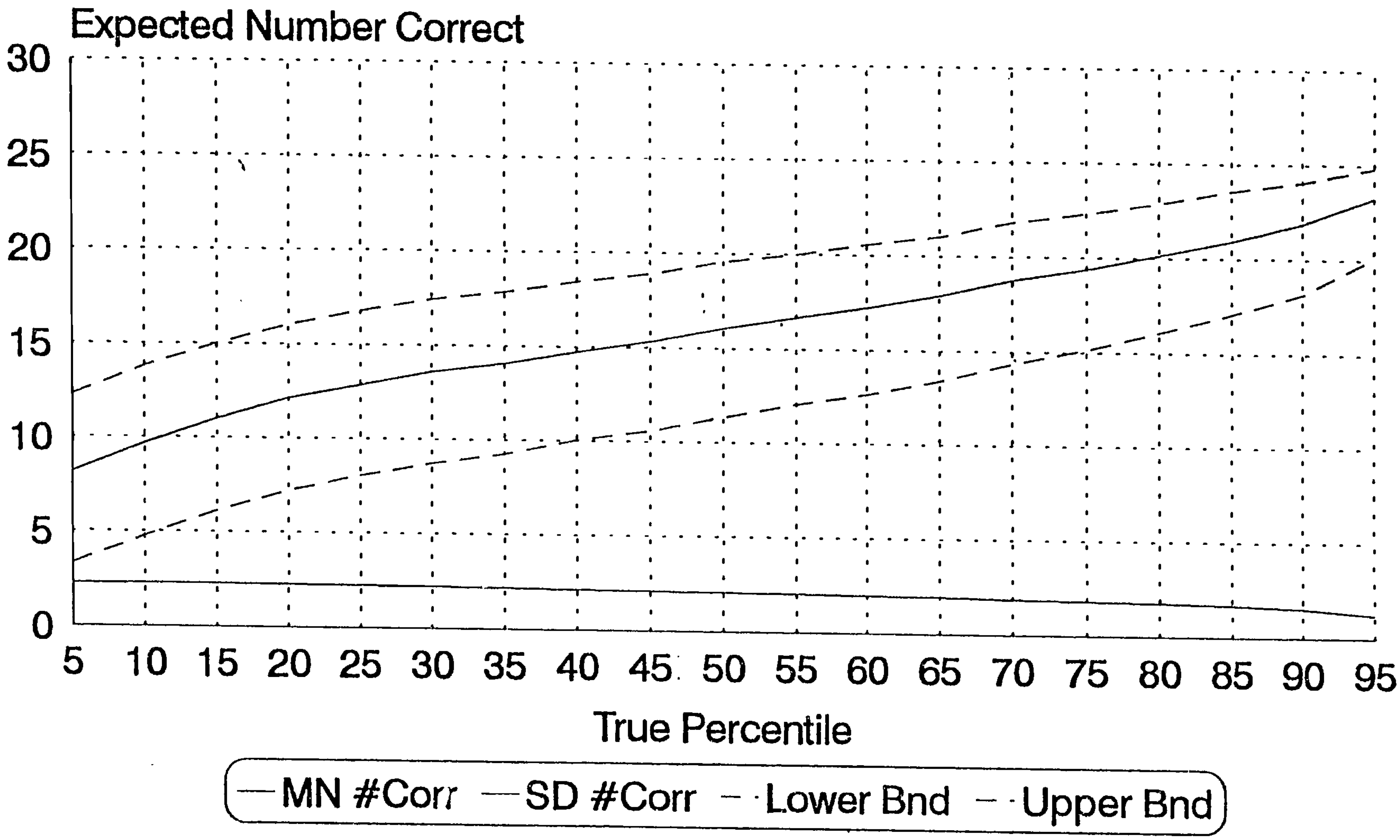
Reference Form WK Subtest



Based on Second 2,000 Case, Self-Weighted Sample

Figure 5a. Expected Number Correct Distribution by Percentile

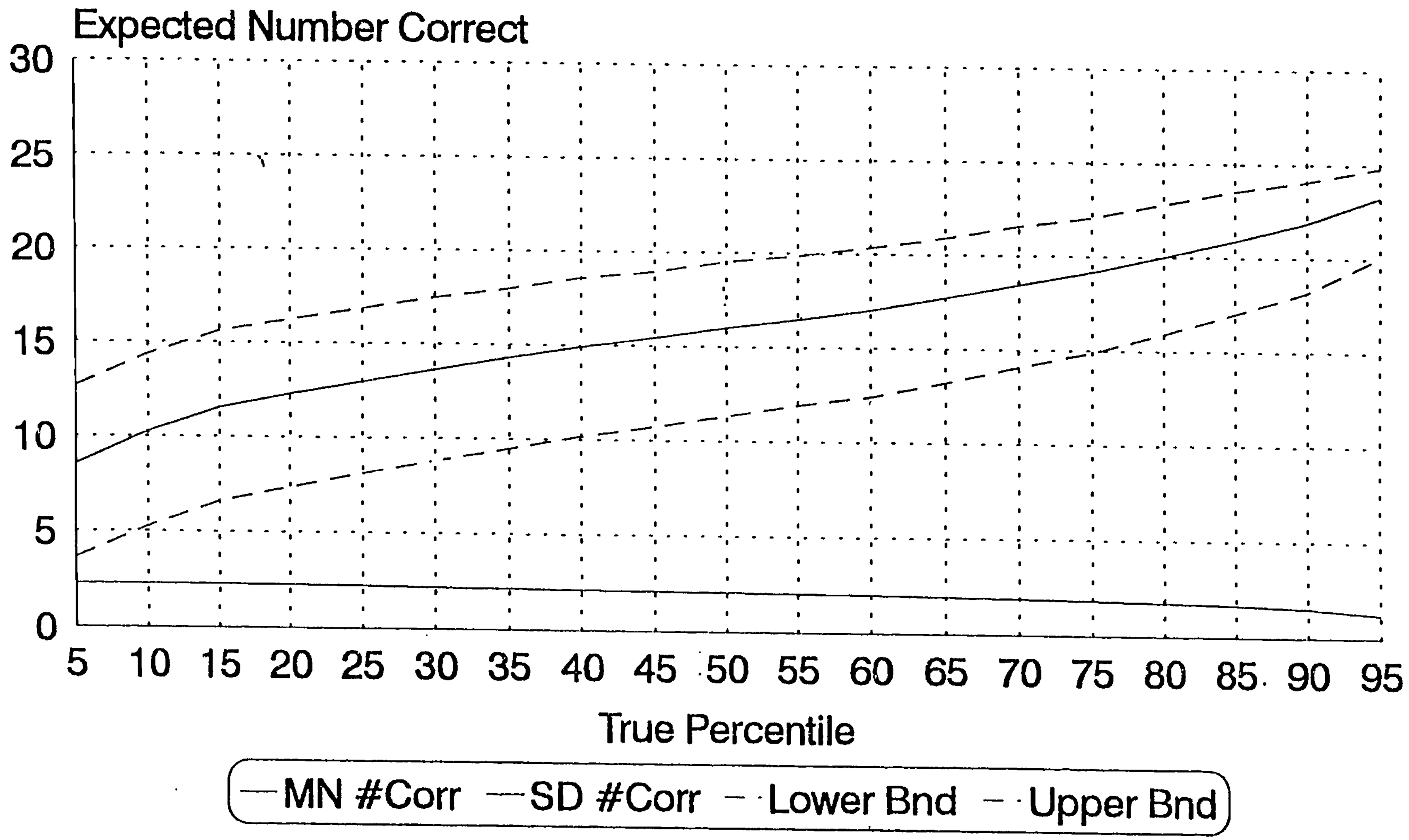
For Reference Form GS Subtest



Based on First 2,000 Case, Self-Weighted Sample

Figure 5b. Expected Number Correct Distribution by Percentile

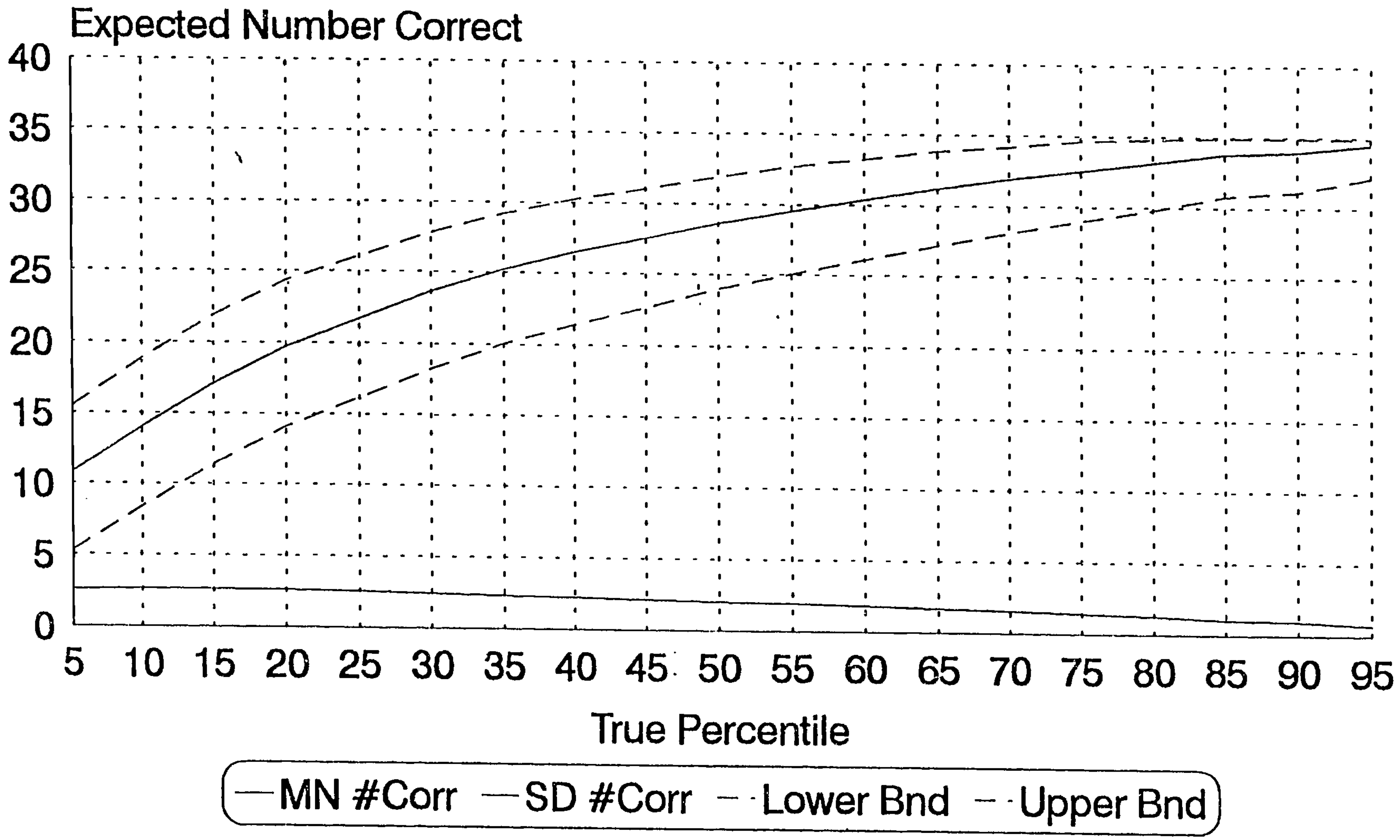
For Reference Form GS Subtest



Based on Second 2,000 Case , Self-Weighted Sample

Figure 6a. Expected Number Correct Distribution by Percentile

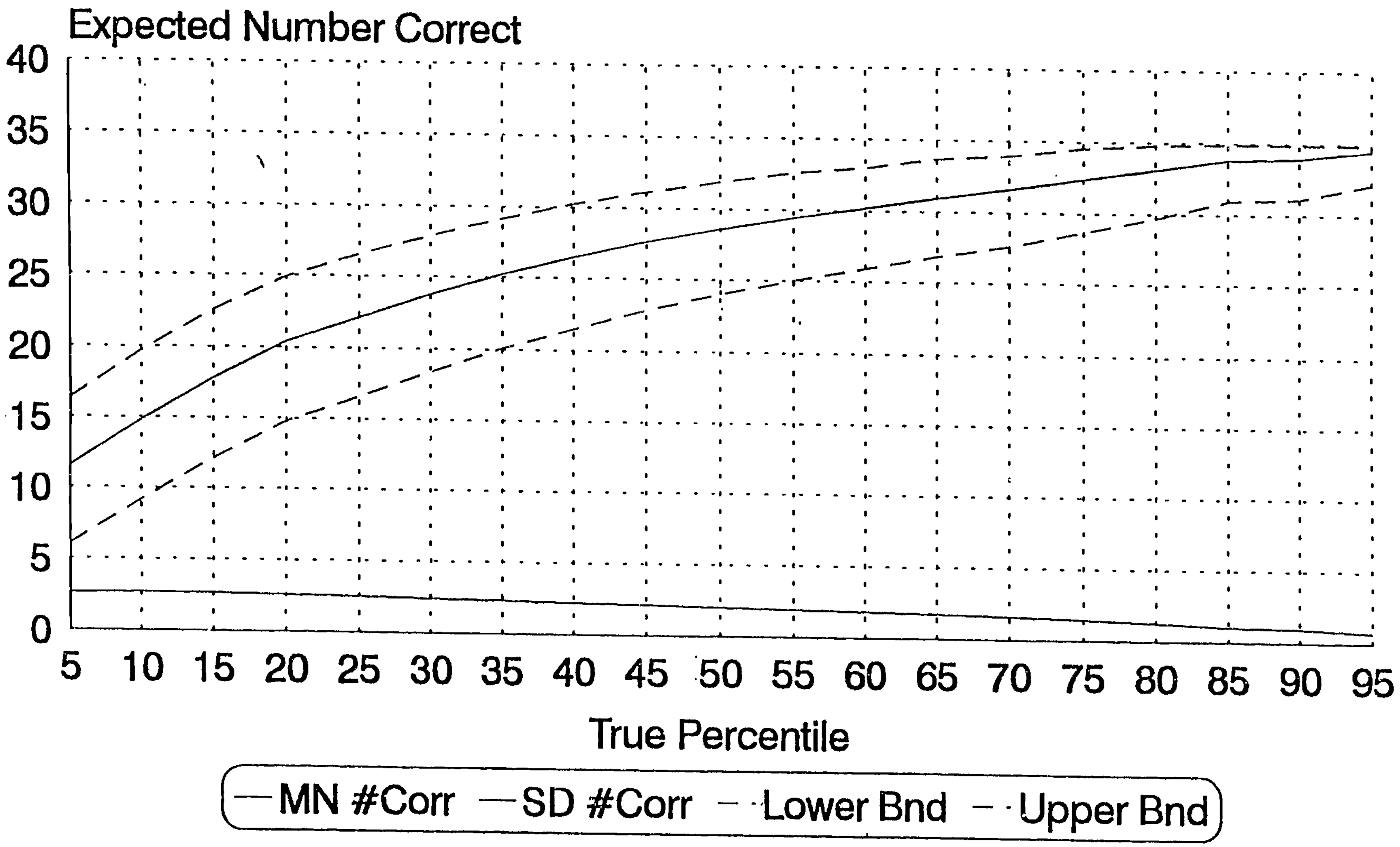
For Reference Form WK Subtest



Based on First 2,000 Case , Self-Weighted Sample

Figure 6b. Expected Number Correct Distribution by Percentile

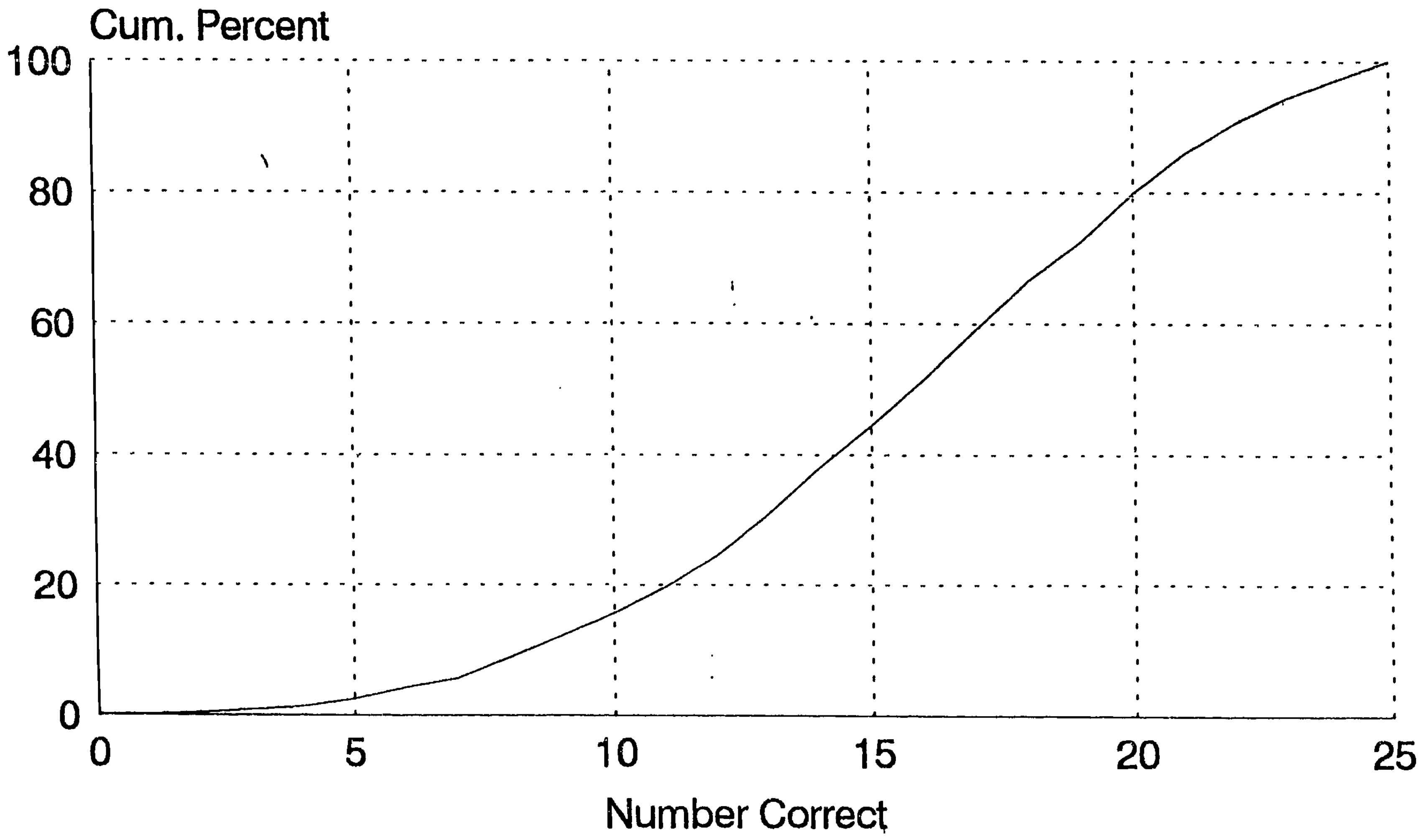
For Reference Form WK Subtest



Based on Second 2,000 Case , Self-Weighted Sample

Figure 7a. Cumulative Distribution of Number Correct Scores

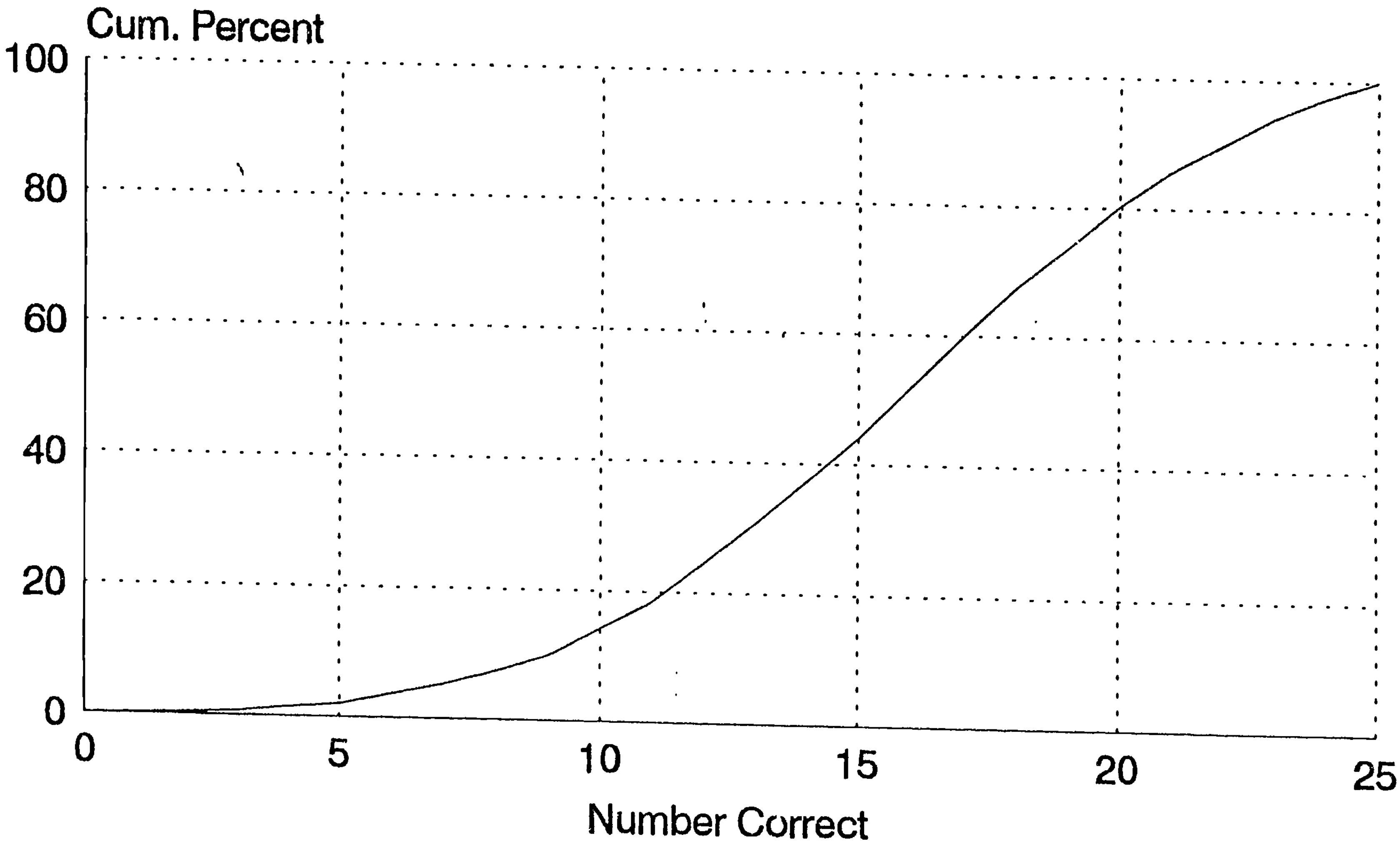
For Reference Form GS Subtest



Based on First 2,000 Case, Self-Weighted Sample

Figure 7b. Cumulative Distribution of Number Correct Scores

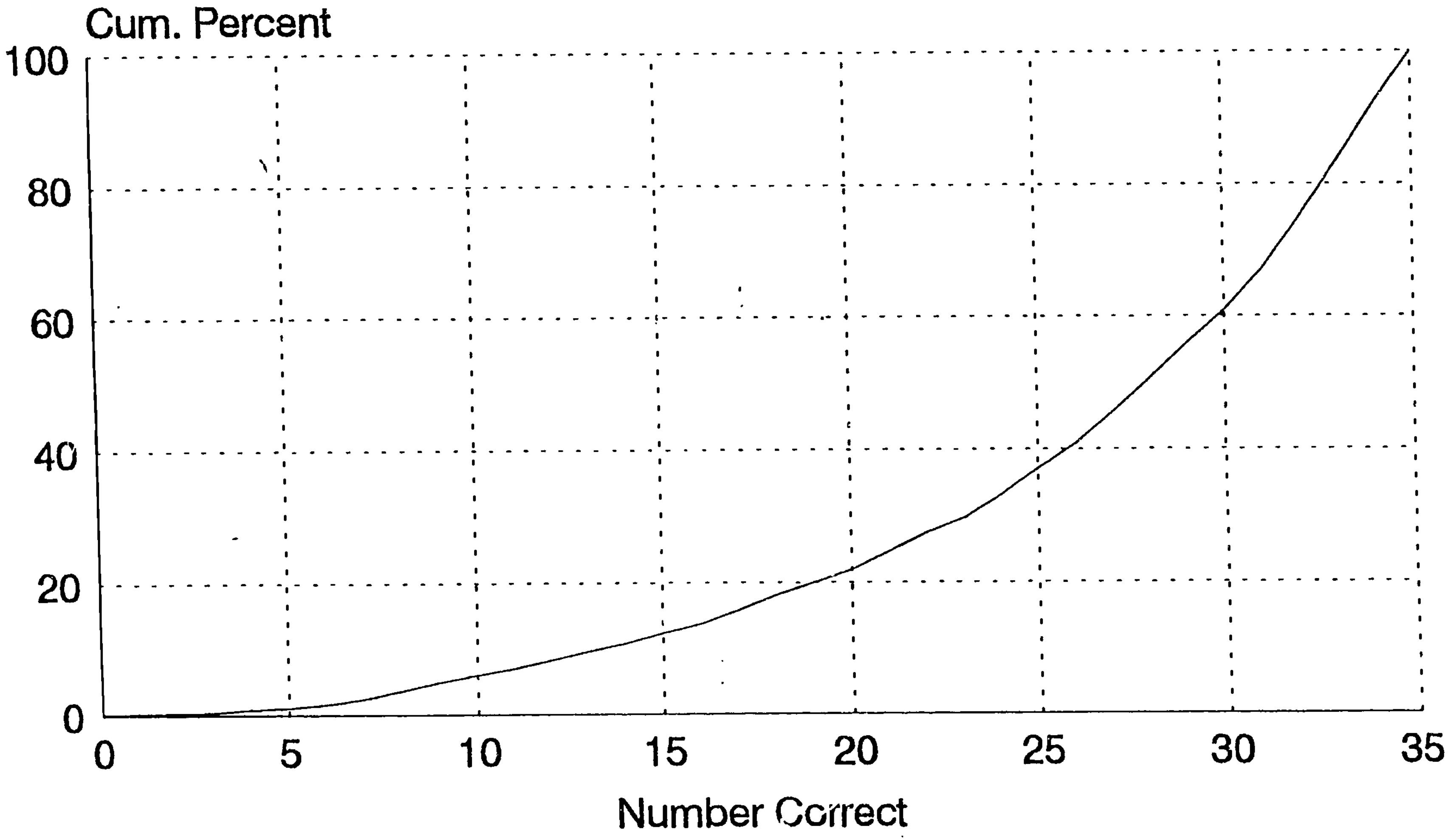
For Reference Form GS Subtest



Based on Second 2,000 Case, Self-Weighted Sample

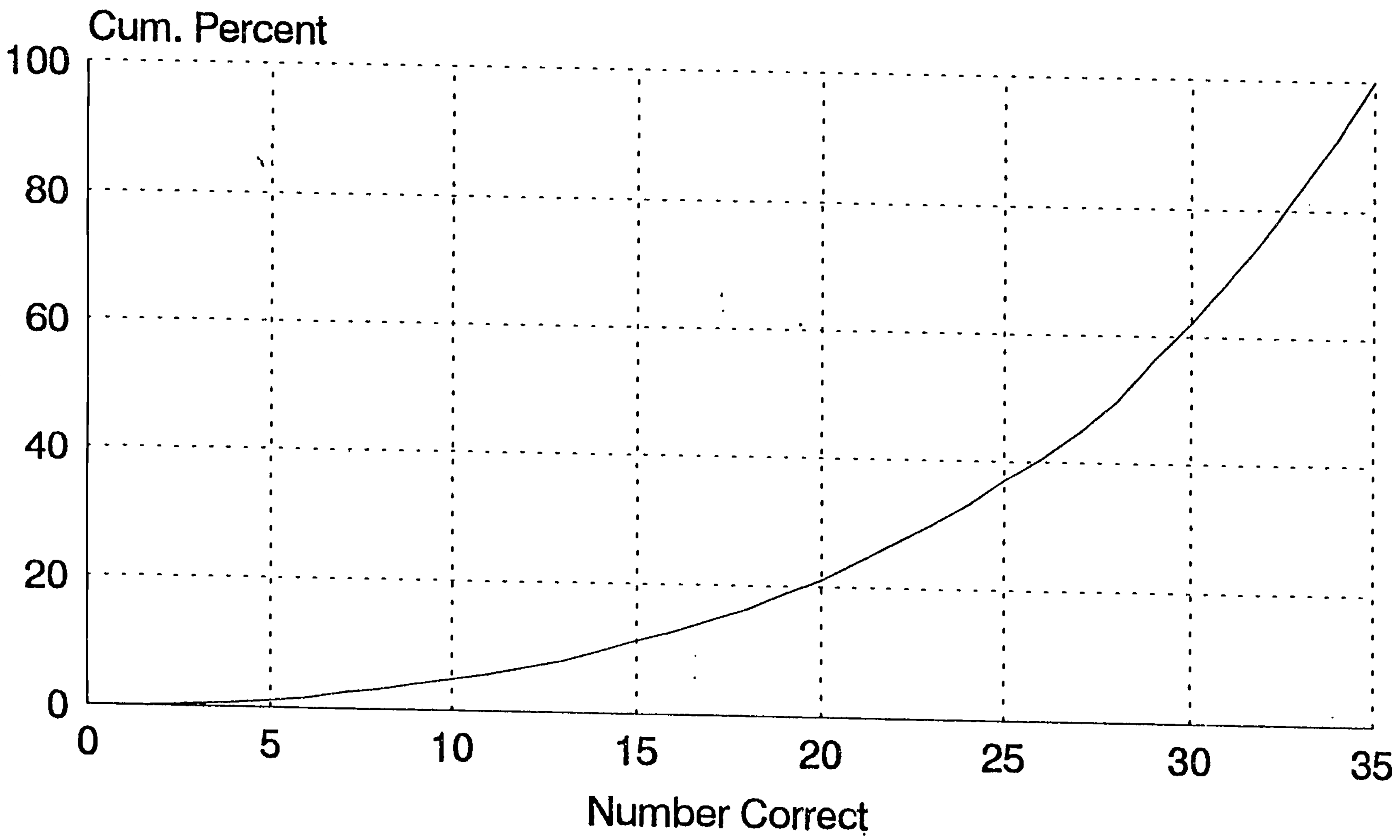
Figure 8a. Cumulative Distribution of Number Correct Scores

For Reference Form WK Subtest



Based on First 2,000 Case, Self-Weighted Sample

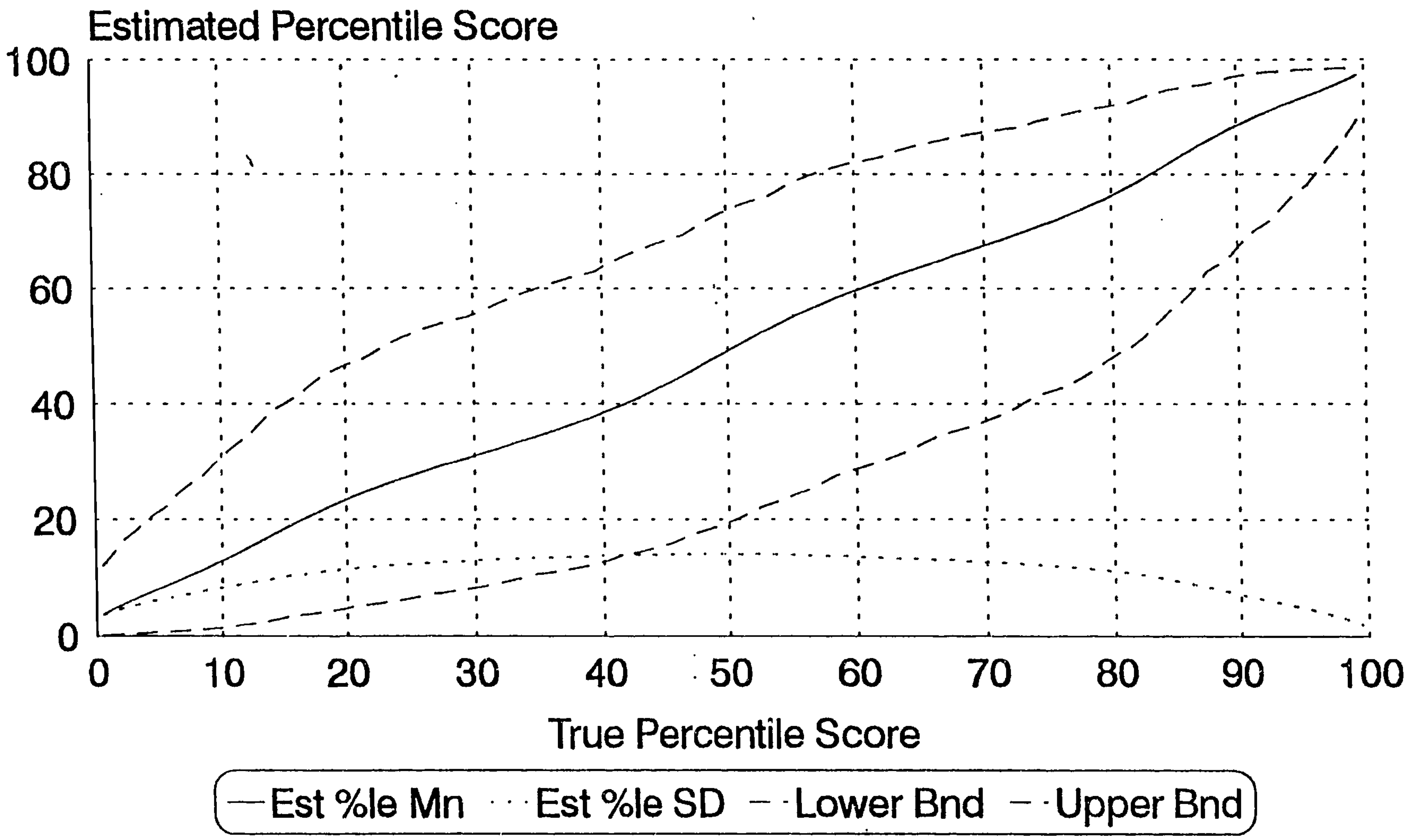
Figure 8b. Cumulative Distribution of Number Correct Scores For Reference Form WK Subtest



Based on Second 2,000 Case, Self-Weighted Sample

Figure 9a. Expected Percentile Score Distribution

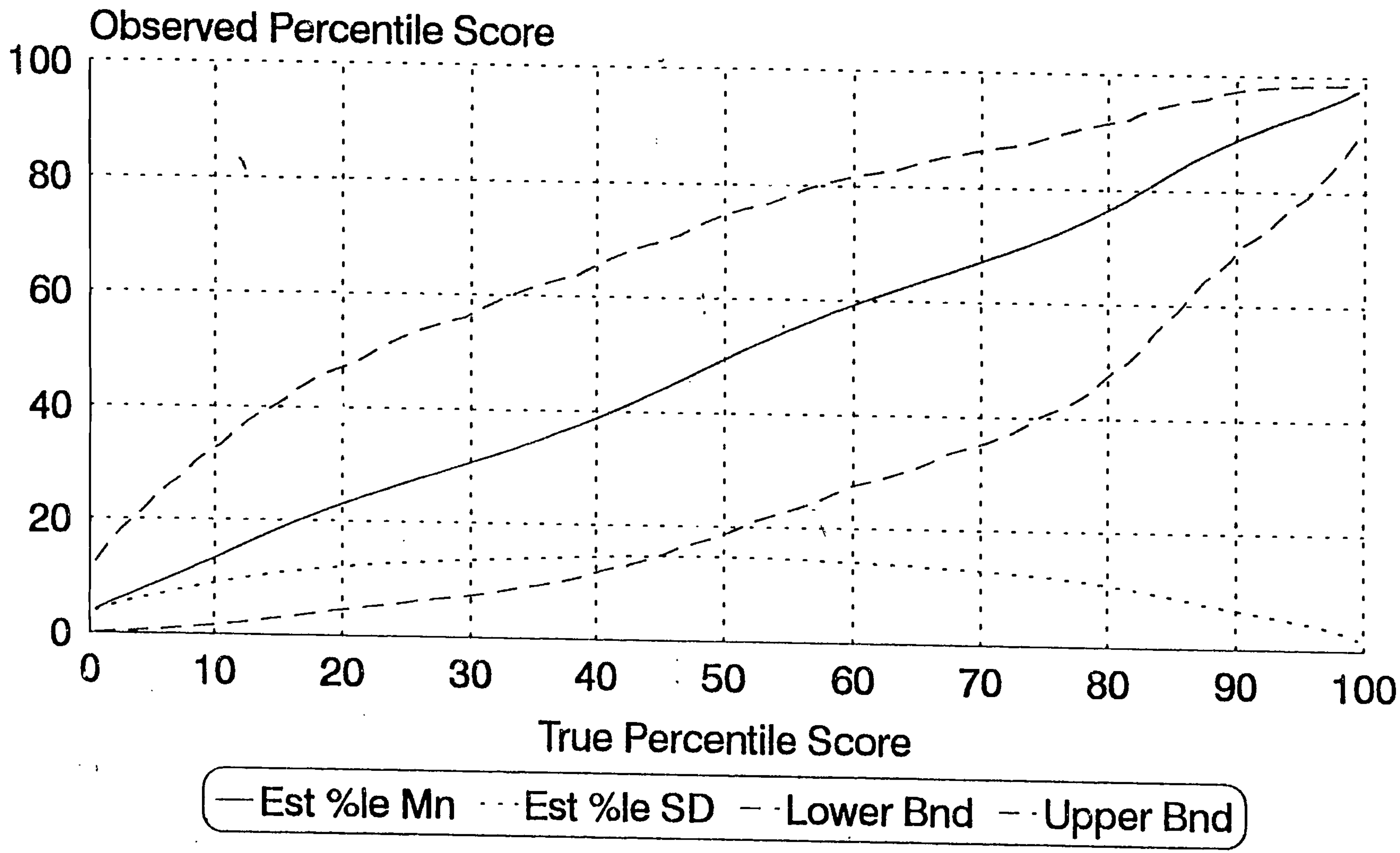
For Reference Form GS Subtest



Based on First 2,000 Case Self-Weighted Sample

Figure 9b. Expected Percentile Score Distribution

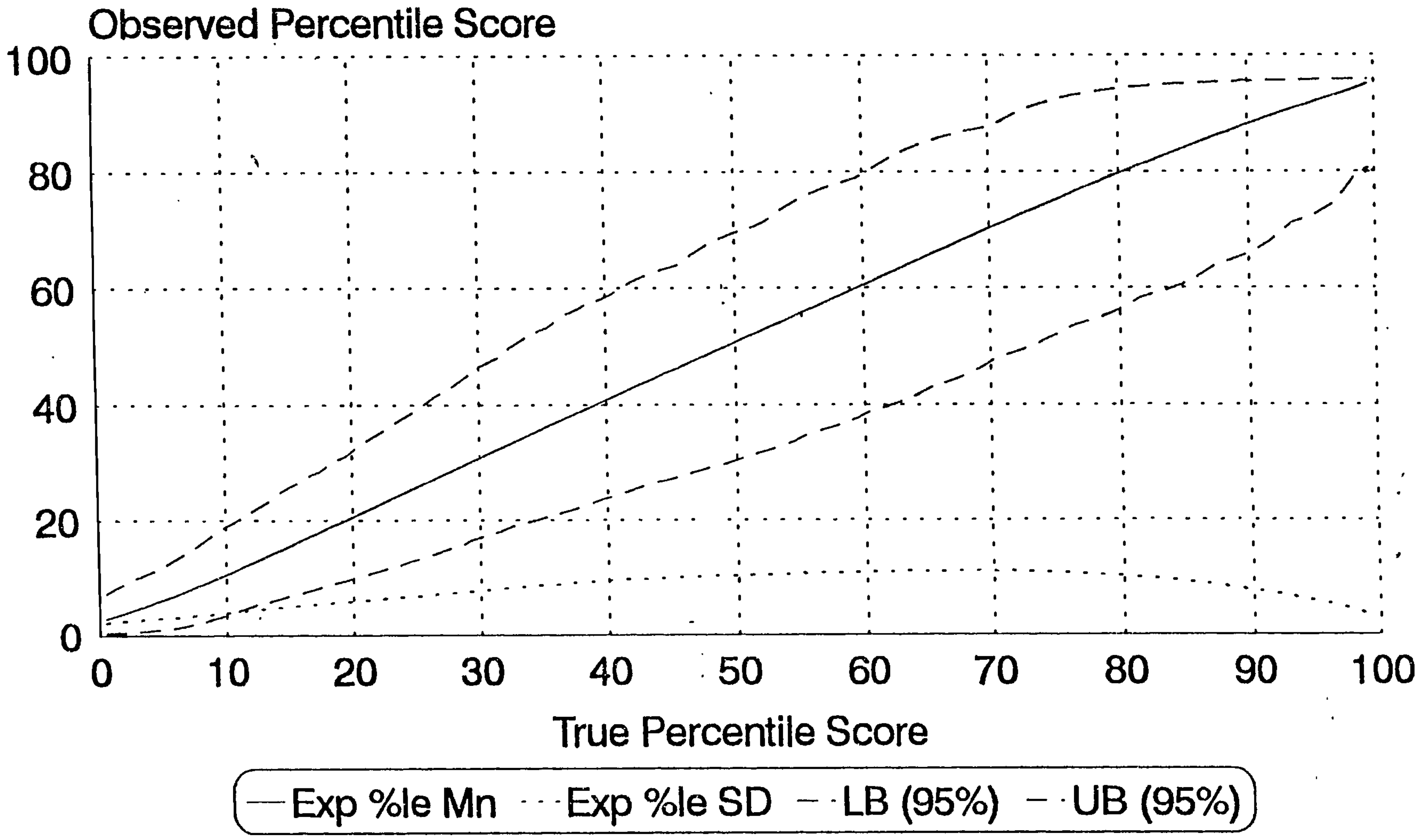
For Reference Form GS Subtest



Based on Second 2,000 Case Self-Weighted Sample

Figure 10a. Expected Percentile Score Distribution

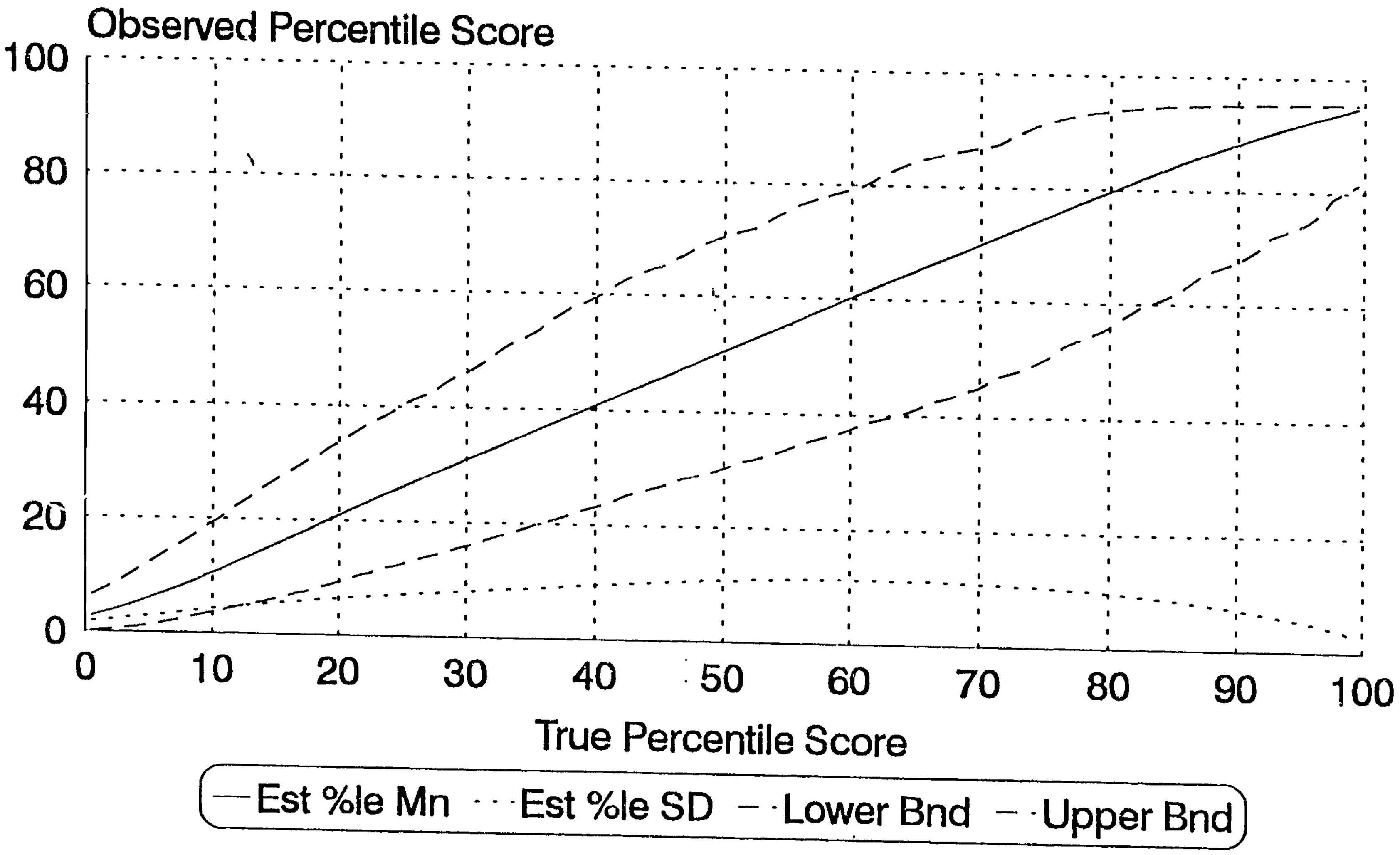
For Reference Form WK Subtest



Based on First 2,000 Case Self-Weighted Sample

Figure 10b. Expected Percentile Score Distribution

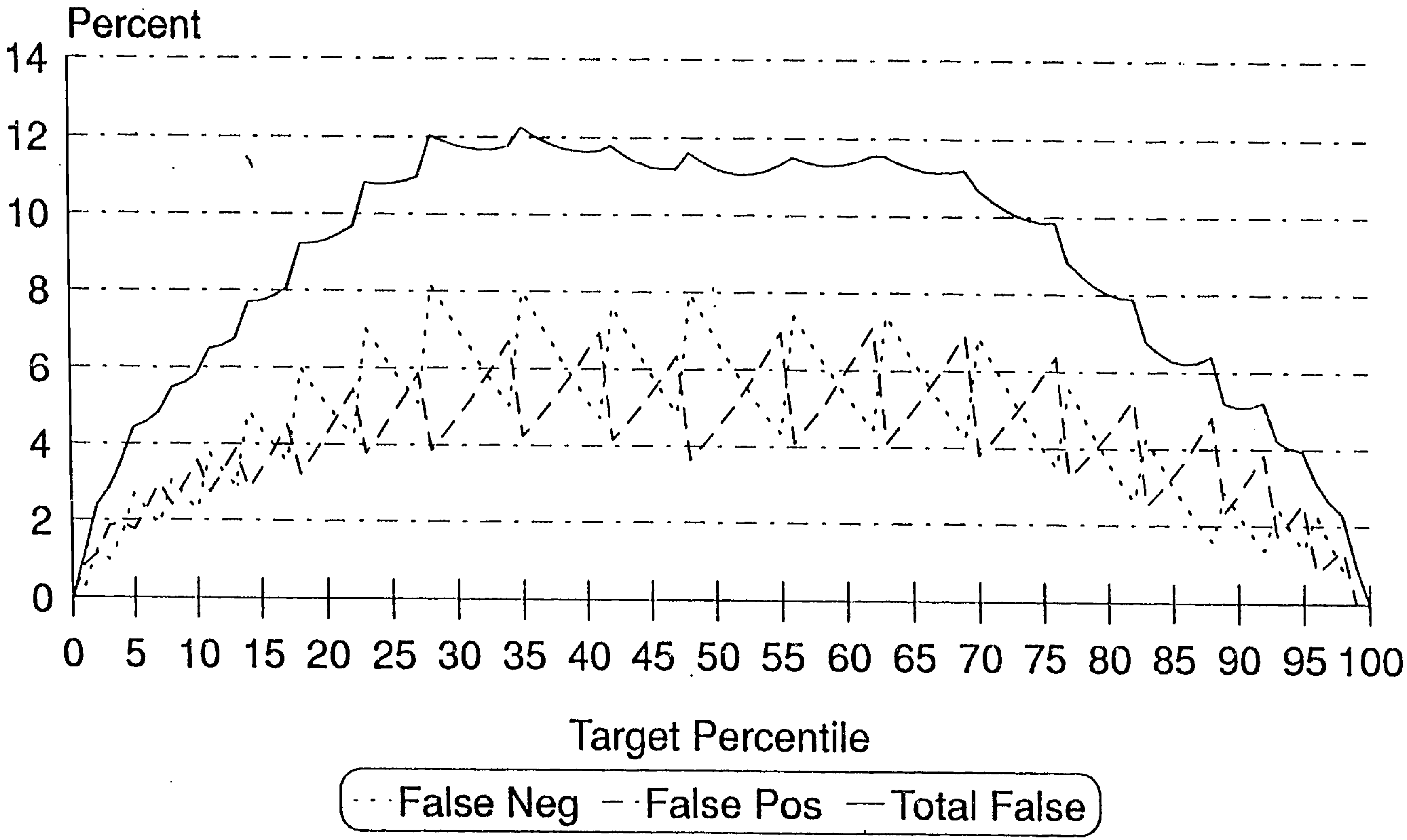
For Reference Form WK Subtest



Based on Second 2,000 Case Self-Weighted Sample

Figure 11a. Classification Error Rates

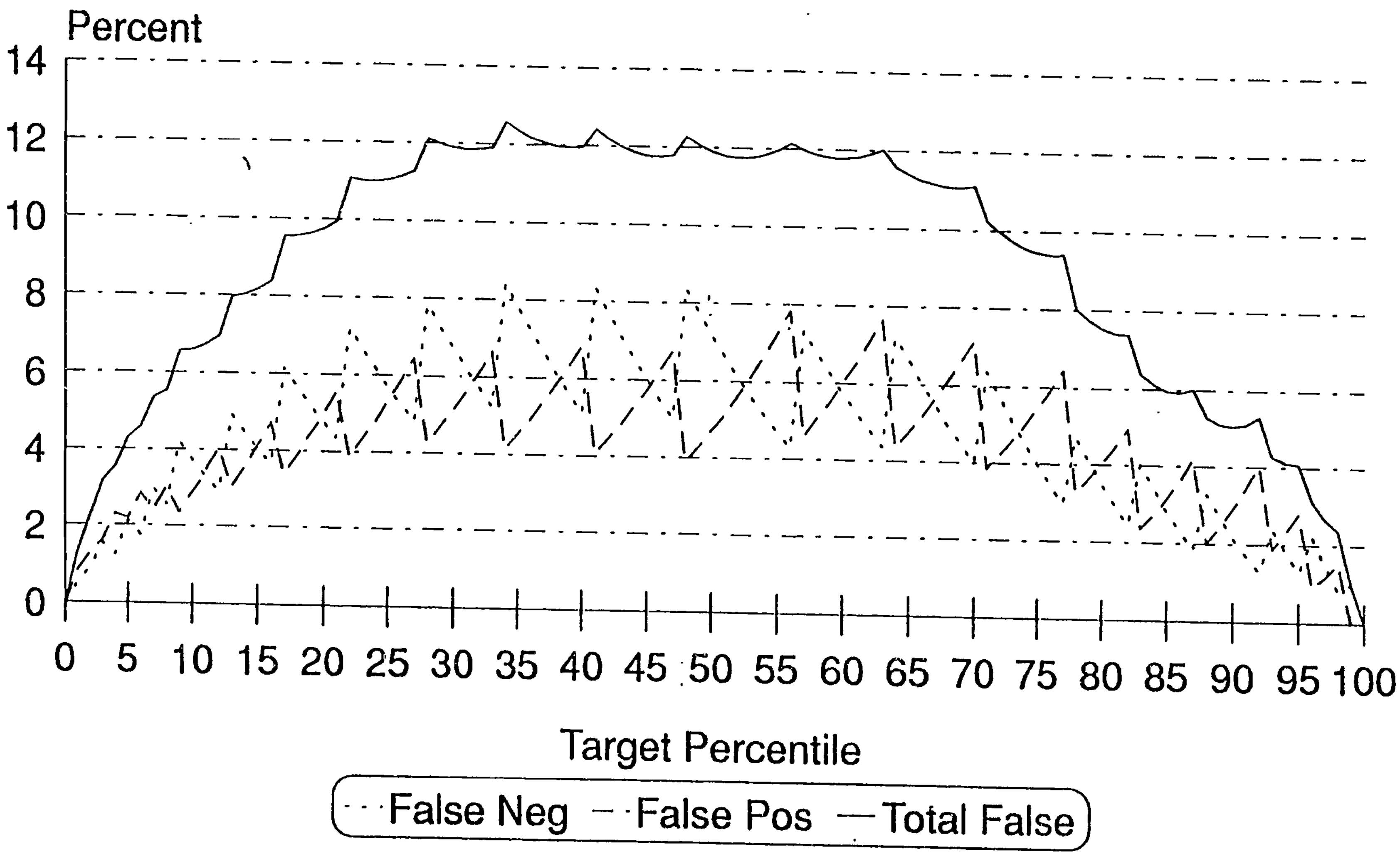
Reference Form GS Subtest



Based on First 2,000 Case Self-Weighted Sample

Figure 11b. Classification Error Rates

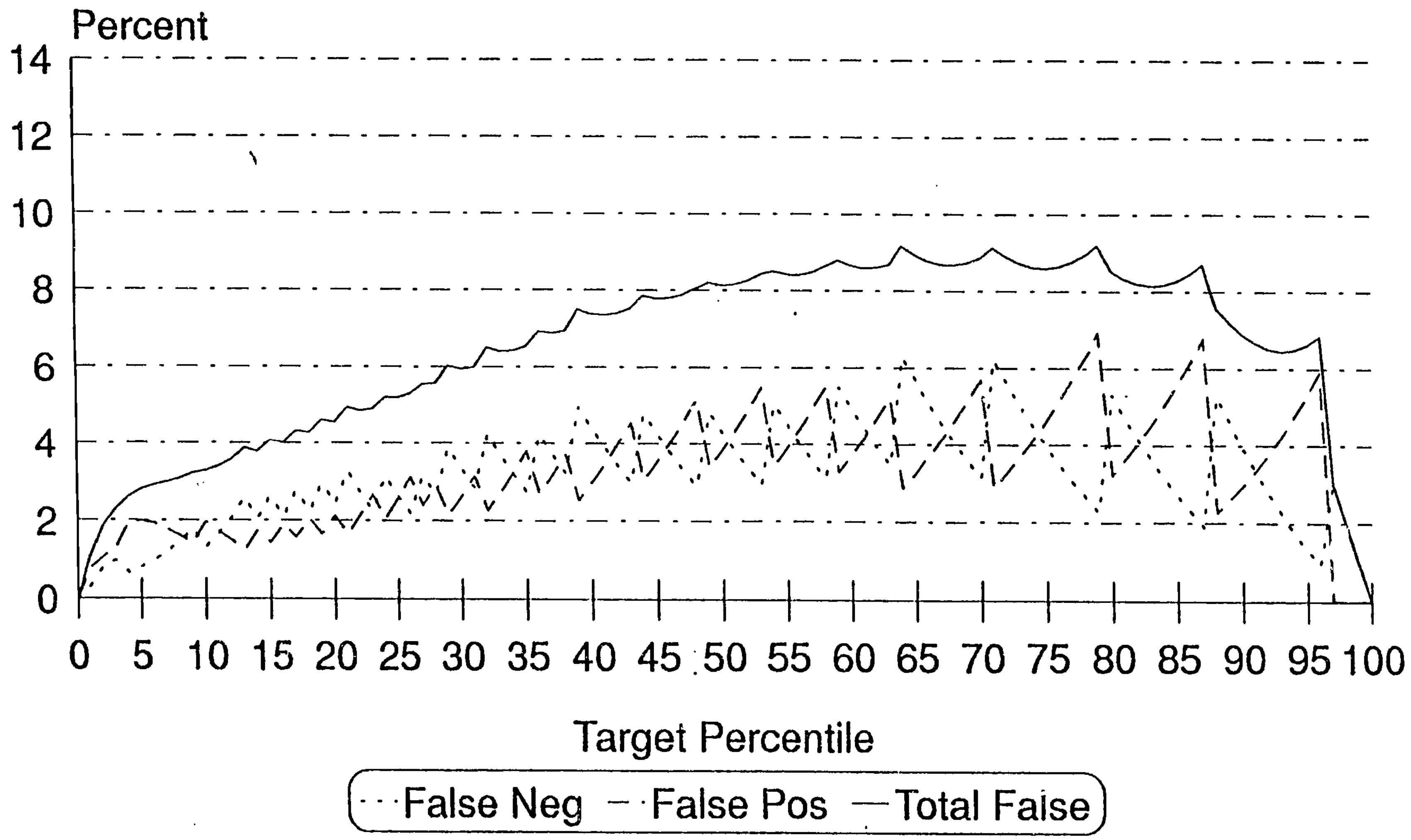
Reference Form GS Subtest



Based on Second 2,000 Case Self-Weighted Sample

Figure 12a. Classification Error Rates

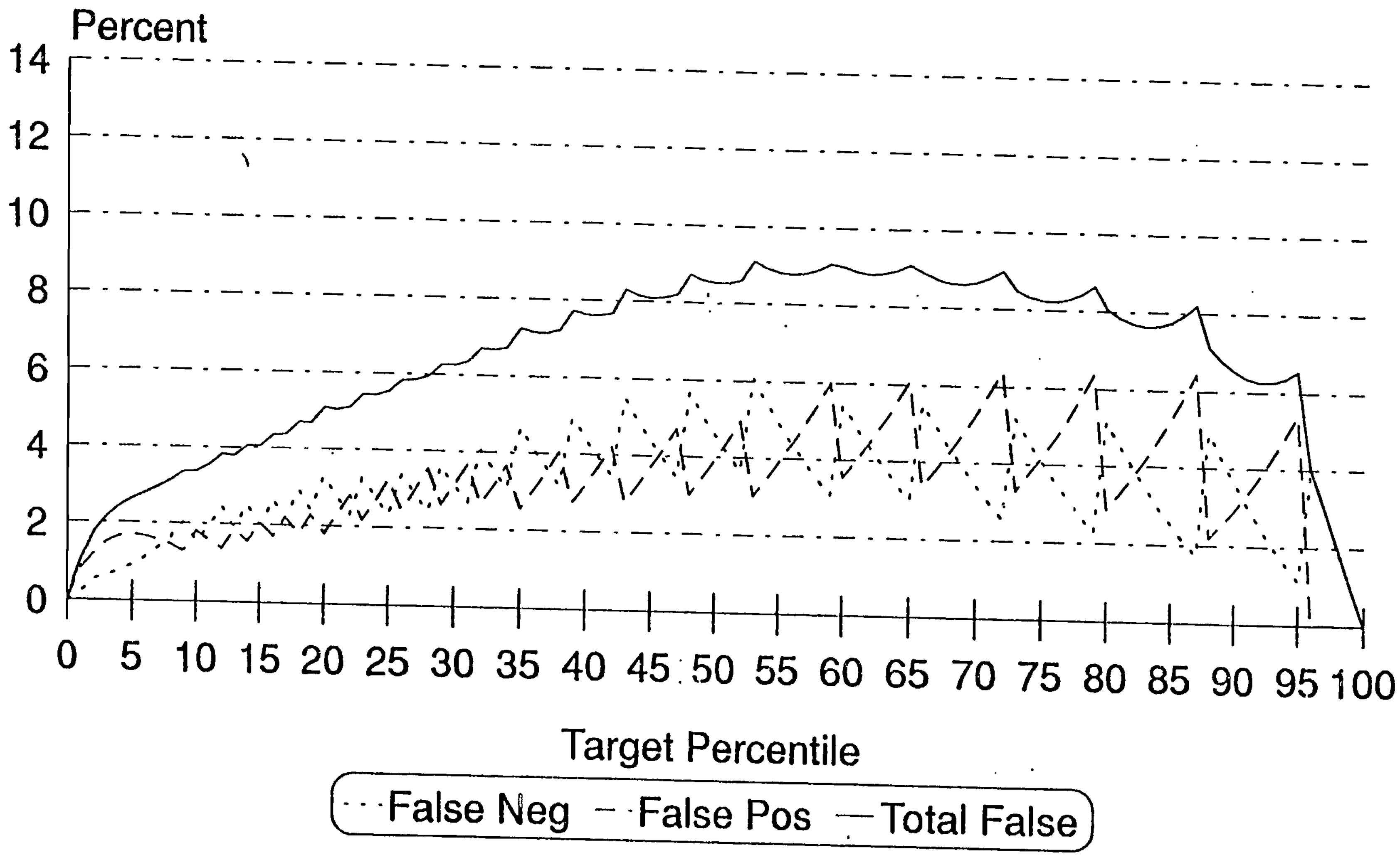
Reference Form WK Subtest



Based on First 2,000 Case Self-Weighted Sample

Figure 12b. Classification Error Rates

Reference Form WK Subtest



Based on Second 2,000 Case Self-Weighted Sample