Research Report

ETS RR–11-28

Measurement Error in Nonparametric Item Response Curve Estimation

Hongwen Guo

Sandip Sinharay

June 2011

Measurement Error in Nonparametric Item Response Curve Estimation

Hongwen Guo and Sandip Sinharay

ETS, Princeton, New Jersey

June 2011

Technical Review Editor: Shelby Haberman

Technical Reviewers: Yi-Hsuan Lee and Frank Rijmen

Abstract

Nonparametric, or kernel, estimation of item response curve (IRC) is a concern theoretically and operationally. Accuracy of this estimation, often used in item analysis in testing programs, is biased when the observed scores are used as the regressor because the observed scores are contaminated by measurement error. In this study, we investigate the deconvolution kernel estimation of IRC, correcting for the measurement error in the regressor variable. Using item response theory (IRT) simulated data and some real data, we compared the traditional kernel estimation and the deconvolution estimation of IRC. Results show that in capturing important features of the IRC, the traditional kernel estimation is comparable to the deconvolution kernel estimation in item analysis.

Key words: IRC, IRT, CTT, measurement error

Acknowledgments

This paper was presented at the annual meeting of the National Council on Measurement in Education (NCME), April 29-May 3, 2010, Denver, CO.

1 Overview

Nonparametric item response theory (NIRT) uses nonparametric regression techniques extensively. (See Douglas & Cohen, 2001; Lee, 2007; Meijer, 2004; and Sijtsma, 1998.) Characterized by a nonparametric function of the latent trait, NIRT differs from parametric item response theory (IRT), which is used in Rasch and the two-parameter and three-parameter logisitic (2PL and 3PL) models. Nonparametric estimation has been the focus of many studies Ramsay (1991) and Wand and Jones (1995) described a nonparametric regression method that can estimate item response curves (IRC). This method, based on kernel smoothing (e.g., Silverman, 1986), is implemented in TESTGRAF (Ramsay, 1998). Livingston and Dorans (2004) discussed the use of Ramsay's (1991) method to estimate the response curves for each answer option for a multiple choice item. Nonparametric estimation is often used in classical test theory (CTT), and recently Lee (2007) compared the kernel smoothing method and other regression methods with monotonicity constraints to estimate item characteristic curves (ICC). We focus only on the nonparametic kernel smoothing methods because the monotonicity constraint on the ICC estimation will not help to identify problematic items. Well-written items should reveal a monotonically increasing IRC for the key, as shown in the left panel of Figure 1. The right panel of Figure 1 shows a decreasing IRC for the key, which indicates that the item may be problematic. An increasing IRC for the top scores of the nonkey also indicates a problematic item. Psychometricians, test developers, and clients find plots similar to those shown in Figure 1 helpful because they are easily interpreted. In the kernel smoothing method (Ramsay, 1998), the response variable is the item score (0 or 1 for the dichotomous items) or the proportion of right answers among examinees; the regressor (or independent variable) is the ability or the total true score of the examinee. In practice, however, neither the ability nor the true score is available. In plotting the IRC, the observed score or scaled score is used, especially for testing programs that use observed scores. (See Figures 1 to 8 in Livingston & Dorans, 2004.) The accuracy of the estimated IRC is a concern because the observed scores are contaminated by measurement error. Nonparametric regression in the presence of measurement error has been studied intensively in the area of statistics. Carroll, Maca, and Ruppert (1999) showed that the simple/nave/traditional nonparametric regression estimate was inconsistent. Fan and Troung (1993) proposed a deconvolution kernel regression method and, under difference measurement error distributions, obtained asymptotic results. Delaigel, Fan, and Carroll (2009) extended the deconvolution method to local polynomial regression. These methods produce asymptotically unbiased estimators; the trade-off is that the convergence rate is discouraging. Wand (1998), however, indicated in a detailed analysis that the deconvolution method can perform well for lower levels of measurement error in reasonable sample sizes. (See also Carroll, Ruppert, Stefanski, & Crainiceanu, 2006.) With this information, it is important to examine if correction for measurement error leads to improved nonparametric estimates of IRCs. Would nonparametric regression estimation of IRC with correction for measurement error result in a significant improvement in identifying problematic items? In this study, we discuss the kernel estimation method (Ramsay, 1991) and then introduce the deconvolution estimation method (Fan & Troung, 1993) in section 2. Section 3 discusses applications of the deconvolution kernel regression method. Naive kernel regression is one of the commonly used nonparametric methods in practice (Livingston & Dorans, 2004; Ramsay, 1998), so we provide a comparison between the naive kernel regression and deconvolution kernel regression of the IRC function by using simulated data and operational data. In section 4, we discuss estimating the IRC function in practice. The distribution of measurement error in both CTT and IRT models are addressed in the appendix.

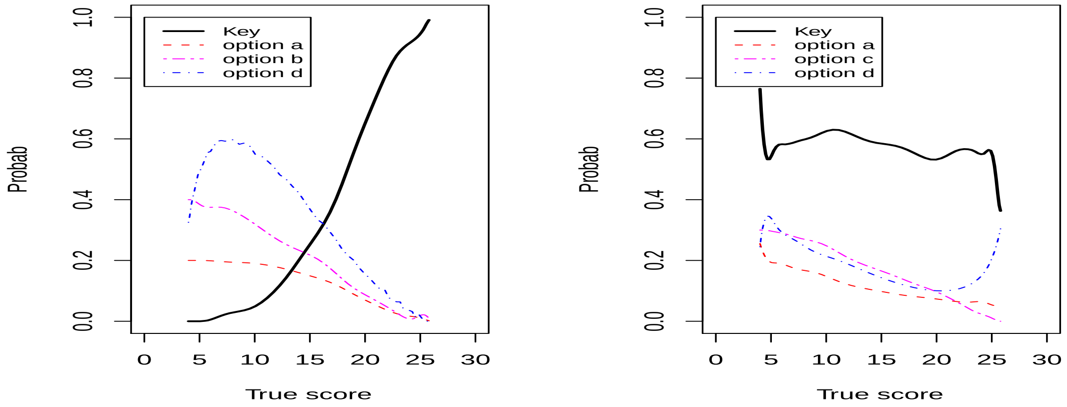

***Figure* 1. Small item response curves (IRCs) of two hypothetical items.**

2 Nonparametric Regression With Measurement Error

Suppose $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independently and identically distributed (i.i.d.) random samples from (X, Y). We are interested in estimating the smooth regression curve $m(x) = E(Y|X = x)$. In the context of educational testing, X_i and Y_i could denote, for examinee

i, the true total score and item score, respectively. When X is observable, at each point x, the (naive) kernel smoothing estimator (Ramsay, 1991; Wand & Jones, 1995) is the weighted average of Y_i:

$$\hat{m}(x) = \sum_{i=1}^{n} K(\frac{x - X_i}{h})Y_i / \sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{1}$$

where $K(\cdot)$ is the kernel function, and h is the bandwidth.

However, sometimes X is not observable. Instead, $(Z_1, Y_1), \cdots, (Z_n, Y_n)$ are observed, where $Z = X + \epsilon$, ϵ is the measurement error and is independent of (X, Y). For example, Z_i could be the observed total score instead of the true total score. The deconvolution method (Fan & Troung, 1993), that can be used to provide a statistically consistent estimator of m based on (Z, Y), is described next.

Notice that $Z = X + \epsilon$, and X and ϵ are independent. Then the probability density function $f_Z(\cdot)$ of Z is a convolution of the two density functions $f_X(\cdot)$ and $f_\epsilon(\cdot)$. That is

$$f_Z(z) = \int f_X(z - x)f_\epsilon(x)dx, \quad \text{i.e.} \quad f_Z = f_X * f_\epsilon.$$

Using the Fourier transformation property (e.g., Stein & Weiss, 1971), one has

$$\mathcal{F}_Z = \mathcal{F}_X + \mathcal{F}_\epsilon,$$

where $\mathcal{F}_X$ is the Fourier transformation of the density function of X. For example,

$$\mathcal{F}_X(t) = \int \exp(-2\pi ixt)f_X(x)dx,$$

where $i = \sqrt{-1}$. Now the convolution problem is simplified as an addition problem. The density function f_X can thus be obtained by an inverse Fourier transform. This is the idea of the deconvolution method. The deconvolution kernel estimator of $m(\cdot)$ is

$$\hat{m}(x) = \sum_{i=1}^{n} K^*(\frac{x - X_i}{h})Y_i / \sum_{i=1}^{n} K^*(\frac{x - X_i}{h}), \tag{2}$$

where

$$K^*(x) = \frac{1}{2\pi} \int \exp(-itx)\frac{\phi_K(t)}{\phi_\epsilon(t/h)}dt \tag{3}$$

and where ϕ_K is the Fourier transform of the kernel function $K(\cdot)$, and ϕ_ϵ is the characteristic function of ϵ:

$$\phi_\epsilon(t) = \int \exp(itx)f_\epsilon(x)dx.$$

Thus, the distribution $f_\epsilon(x)$ of the measurement error should be known in order to use the deconvolution method. See the appendix for some results on distributions of measurement error.

The deconvolution estimation produces an asymptotically unbiased estimator, but the convergence rate of the deconvolution estimator is slower than the naive kernel smoothing estimator (Fan & Troung, 1993).

Fan and Troung (1993) showed that the deconvolution estimator is robust to different choices of kernel functions. Among these kernels, one has the following simple form:

$$K^*(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) \Big[1 - \frac{\sigma_\epsilon^2}{2h^2}(x^2 - 1) \Big],$$

which will be used in the IRC estimation later. This K^* is different from a regular kernel K. Figure 2 displays how different K^* is from K for different standard errors of measurement (SEMs) and bandwidths when K is a normal density ψ.

From Figure 2, we can observe that when the standard deviation of error σ_ϵ (denoted as SEM in the plots) is very small, K^* and ψ are hardly distinguishable for a wide range of bandwidths h. As σ_ϵ increases, K^* deviates more from ψ. But h has the opposite effect on K^*; as h decreases, K^* deviates more from ψ.

3 Applications

3.1 Simulated Data

We use the 2PL IRT model to simulate data in order to have a true item characteristic function (IRF) to compare with the nonparametric estimations of IRC. For a 2PL model, the IRF for item j is given by (A6). Given a test form with $(\alpha_j, \beta_j), j = 1, \cdots, J$, the true score X given θ is

$$X(\theta) = \sum_{j=1}^{J} P(\theta; \alpha, \beta) = \sum_{j=1}^{J} \frac{e^{\alpha_j \theta - \beta_j}}{1 + e^{\alpha_j \theta - \beta_j}},$$

which is a monotonic function of θ. Thus there exists a one-to-one relationship between the true score $X(\theta)$ and the ability parameter θ. The plot of $P(\theta; j)$ against $X(\theta)$ is our criterion in the comparison of different ways of nonparametric IRC estimation of item j.

In the simulation, test lengths are 20, 40, and 80 for short, medium and long tests. Sample sizes are 100, 500, 1,000, and 5,000 for small, medium, large, and very large samples The ability

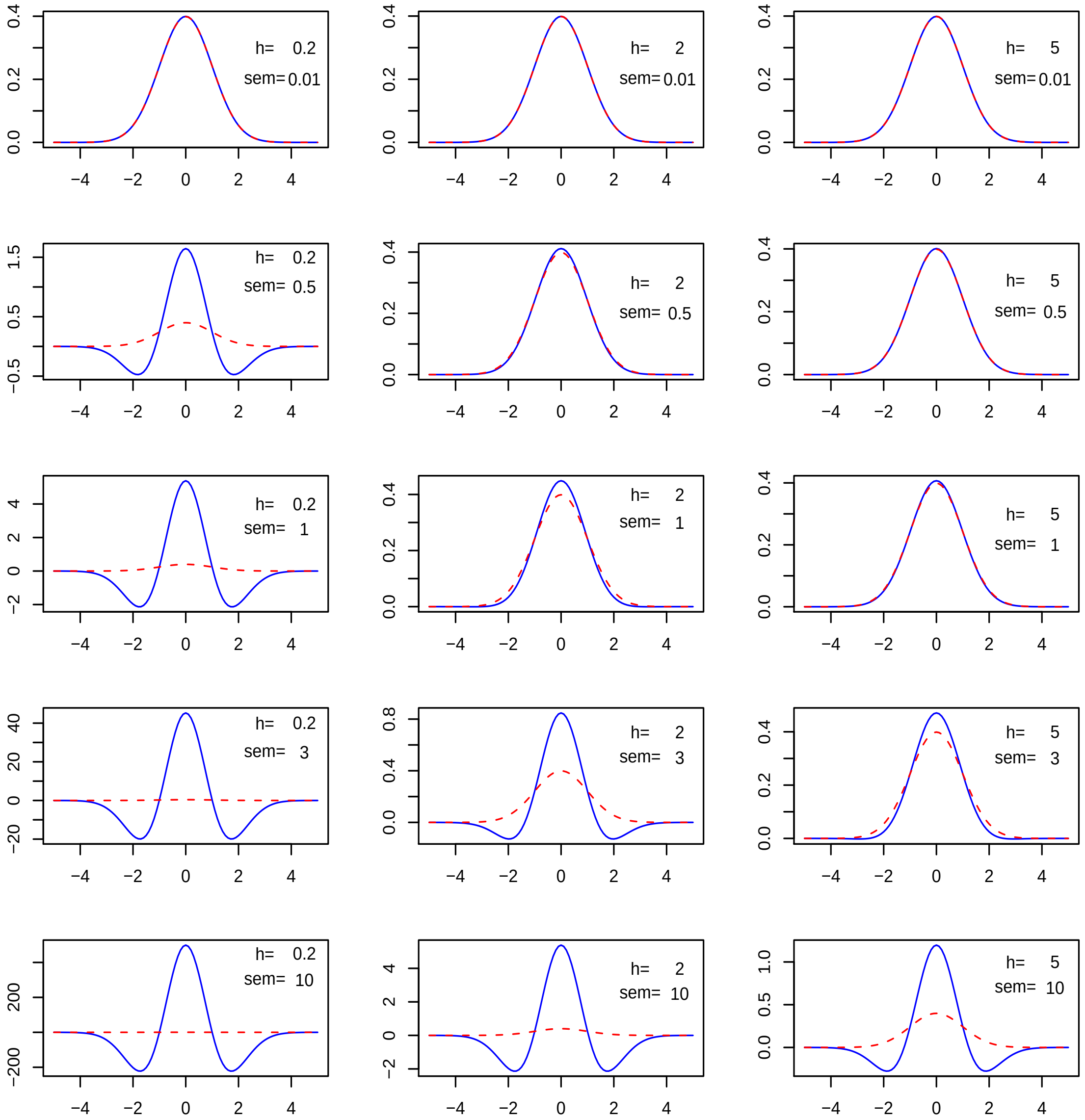

***Figure* 2. A comparison of the deconvolution kernel $K^*(x)$ (solid line) and the normal density $\psi(x)$ (dashed line) for different bandwidth h and standard deviation σ_ϵ of error. In each plot, the x-axis indicates the independent variable x, and the y-axis indicates the dependent variable $K^*(x)$ or $\psi(x)$. Notice that the y-axis scales are different in the plots.**

follows a normal distribution $N(0, 1)$; the item difficulty follows the same distribution as the ability distribution; the item discrimination follows a normal distribution $N(1, .25)$.

Calculation of $\hat{m}(x)$ requires specification of a bandwidth h. For the naive kernel smoothing, a popular approach is to make an asymptotic expansion of MISE mean integrated squared error (MISE)

$$MISE = E\Big(\int [\hat{m}(x) - m(x)]^2 f(x)dx\Big),$$

where $f(x)$ is the density function of X. The optimal bandwidth is the one that minimizes MISE (Ruppert, Sheather, & Wand, 1995):

$$h_{\mathrm{MISE}} = \Big[\frac{R(K)}{\mu_2(K)^2 \int m^{(2)}(x)^2 f(x)dx}\Big]^{1/5} n^{-1/5}, \tag{4}$$

where $R(K) = \int K^2(x)dx$, $\mu_2(K) = \int x^2 K(x)dx$, and $m^{(2)}(x)$ is the second derivative function of $m(x)$. Replacing the unknown integrals by estimators gives the plug-in bandwidth. In the following estimation, the bandwidth for the naive kernel smoothing is h_{MISE}. Notice that this h_{MISE} helps to produce a smooth regression curve.

Let the root sum squared error (RSE) of the nonparametric estimate of the IRC be

$$\mathrm{RSE} = \sqrt{\sum_{i=1}^{n} (\hat{m}(x_i) - m(x_i))^2},$$

where $\hat{m}(x)$ is the estimated function and $m(x)$ is the true function. Since there is no available optimal bandwidth formula for the deconvolution estimation, we experimented with different bandwidths and found that the deconvolution estimation with a half of h_{MISE} produces a IRC with a minimum RSE or even smaller one. Therefore, in the following estimation, the bandwidth for the deconvolution method is chosen to be $h_{\mathrm{MISE}}/2$.

Table 1 compares the RSE of the naive estimators (RSE.n) and RSE of the deconvolution estimator (RSE.d) of IRC for one item under different sample sizes, test lengths, and ability variance. Notice that the bandwidth is chosen as in (4), which minimizes the MISE for the naive kernel estimation, not the deconvolution estimation. The RSE.d could have been slightly improved had we adjusted the bandwidth for individual item. However, the improvement was found to be negligible in our analysis.

Figures 3 to 5 compare the deconvolution estimator and the naive kernel estimator for three simulated data sets. The measurement error is assumed to have normal distribution

$N(0, \sigma_\epsilon)$, where σ_ϵ is estimated by SEM.th in (A7) in the appendix. It can be observed that the deconvolution estimators and the naive estimators behave similarly with respect to the criterion for all concerned test lengths, numbers of examinees, and standard deviations of ability distributions. Figure 6 displays the IRC estimates with different ability standard deviations.

3.2 Real Data Examples

We now compare the two nonparametric IRC estimators using a data set from an operational test. The test has 147 multiple-choice items. Its SEM of 6.9, which is calculated from (A2) in the appendix, is used as σ_ϵ in the calculation of the deconvolution estimator. Also, the error is assumed to follow a normal distribution. Figure 7 displays the IRC plots of 12 items on the test. For comparison, we also included Ramsay's estimator (Ramsay, 1991) in the plots. All the estimators of IRC capture the characteristics (shape, monotonicity, etc.) of the IRC in the same way. Note that each panel in Figure 7 shows the nonparametric estimate corresponding to only the key of multiple-choice items. It is possible to compare plots like those in Figure 1 by computing the nonparametric estimates of the nonkey answer options using both the deconvolution and

Table 1

Comparison of RSE.n and RSE.d for One Item

	σ_θ	.1			1			5		
	J	20	40	80	20	40	80	20	40	80
n = 100	RSE.n	0.36	0.53	0.75	1.19	1.46	0.72	0.65	1.24	0.42
	RSE.d	0.39	0.72	0.87	1.02	1.38	0.66	0.59	1.25	0.39
n = 500	RSE.n	0.54	0.40	0.27	0.43	1.60	1.06	1.38	1.38	0.77
	RSE.d	0.79	0.38	1.31	0.47	1.42	1.01	1.23	1.27	0.81
n =1,000	RSE.n	0.39	0.53	0.91	1.50	1.35	1.28	1.16	0.49	0.92
	RSE.d	0.30	1.20	1.80	0.91	1.05	1.11	1.75	1.07	0.88
n = 5,000	RSD.n	1.18	0.23	0.64	2.12	2.25	4.03	1.32	1.31	1.14
	RSD.d	1.08	0.80	0.40	0.91	1.70	4.00	2.56	0.93	1.10

Note. RSE.d = root sum squared error (RSE) of the deconvolution estimator, RSE.n = RSE of the naive estimators.

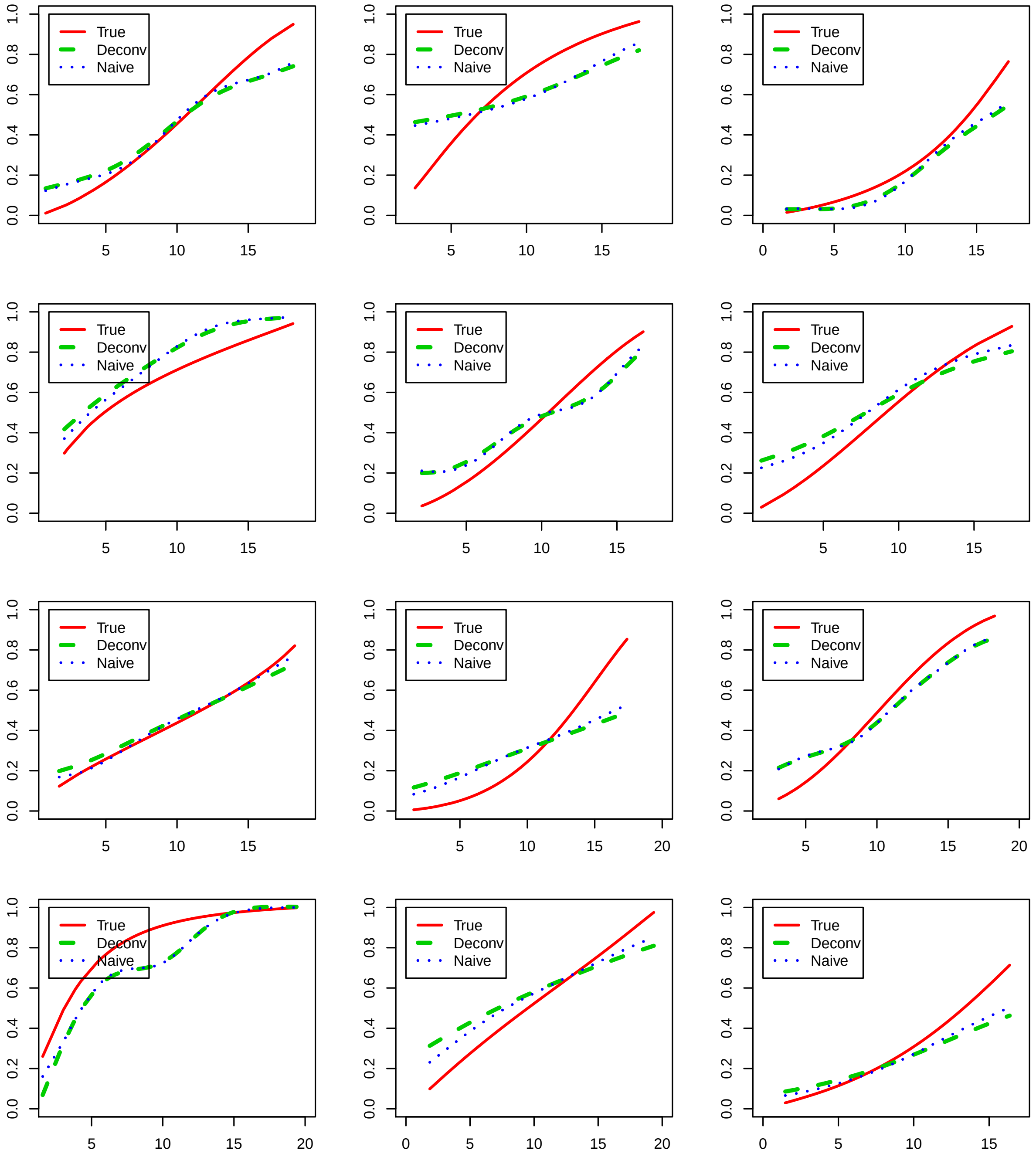

***Figure* 3. Item response curve (IRC) plots of several items in a test of length 20 and sample size 100. In each plot, the x-axis indicates the true score, and the y-axis is the probability of answering the item right for a true score x.**

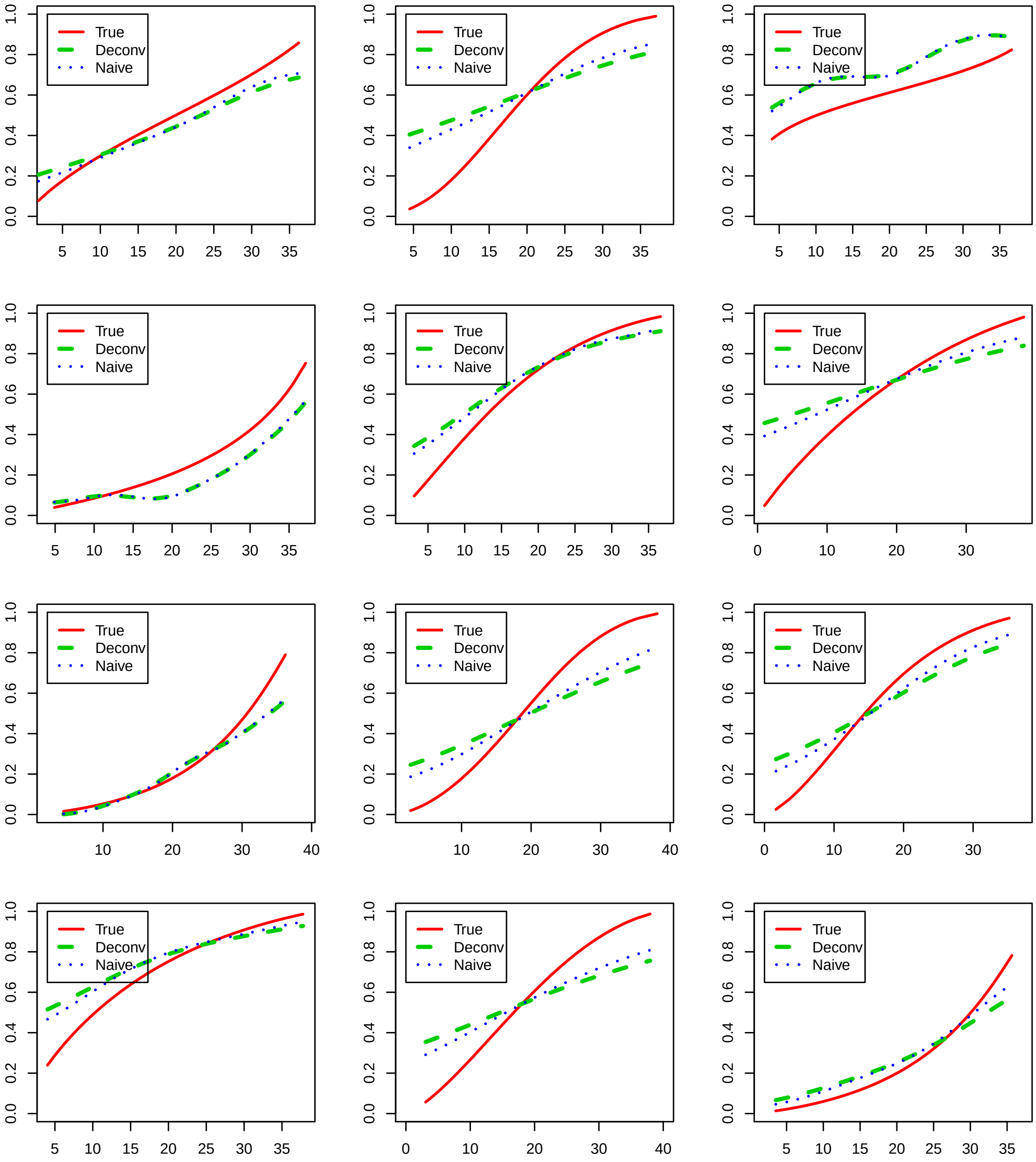

***Figure* 4. Item response curves (IRC) plots of several items in a test of length 40 and a sample size of 500. In each plot, the x-axis indicates the true score, and the y-axis is the probability of answering the item right for a true score of x.**

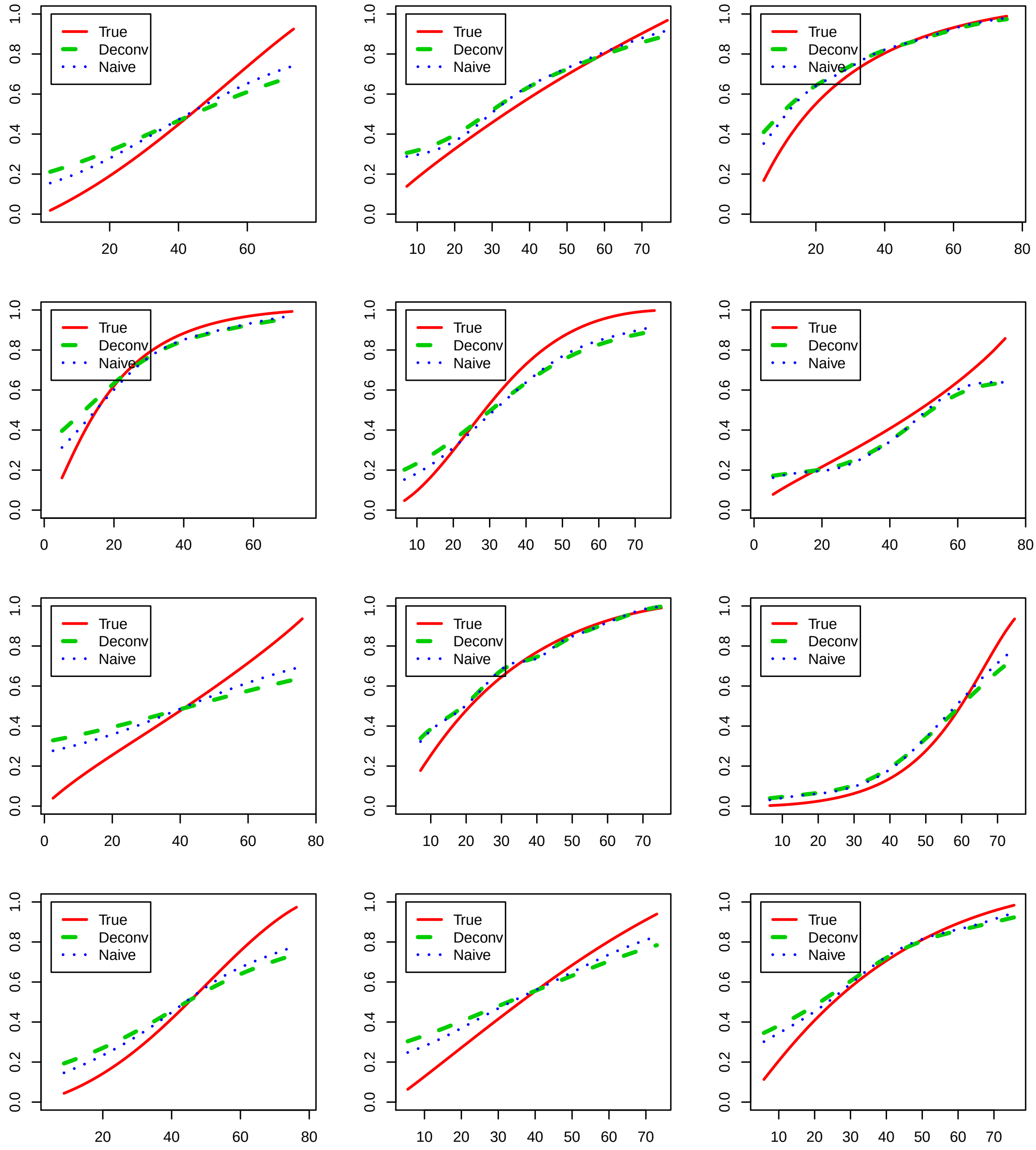

***Figure* 5. Item response curve (IRC) plots of several items in a test of length 80 and a sample size of 1,000. In each plot, the x-axis indicates the true score, and the y-axis is the probability of answering the item right for a true score of x.**

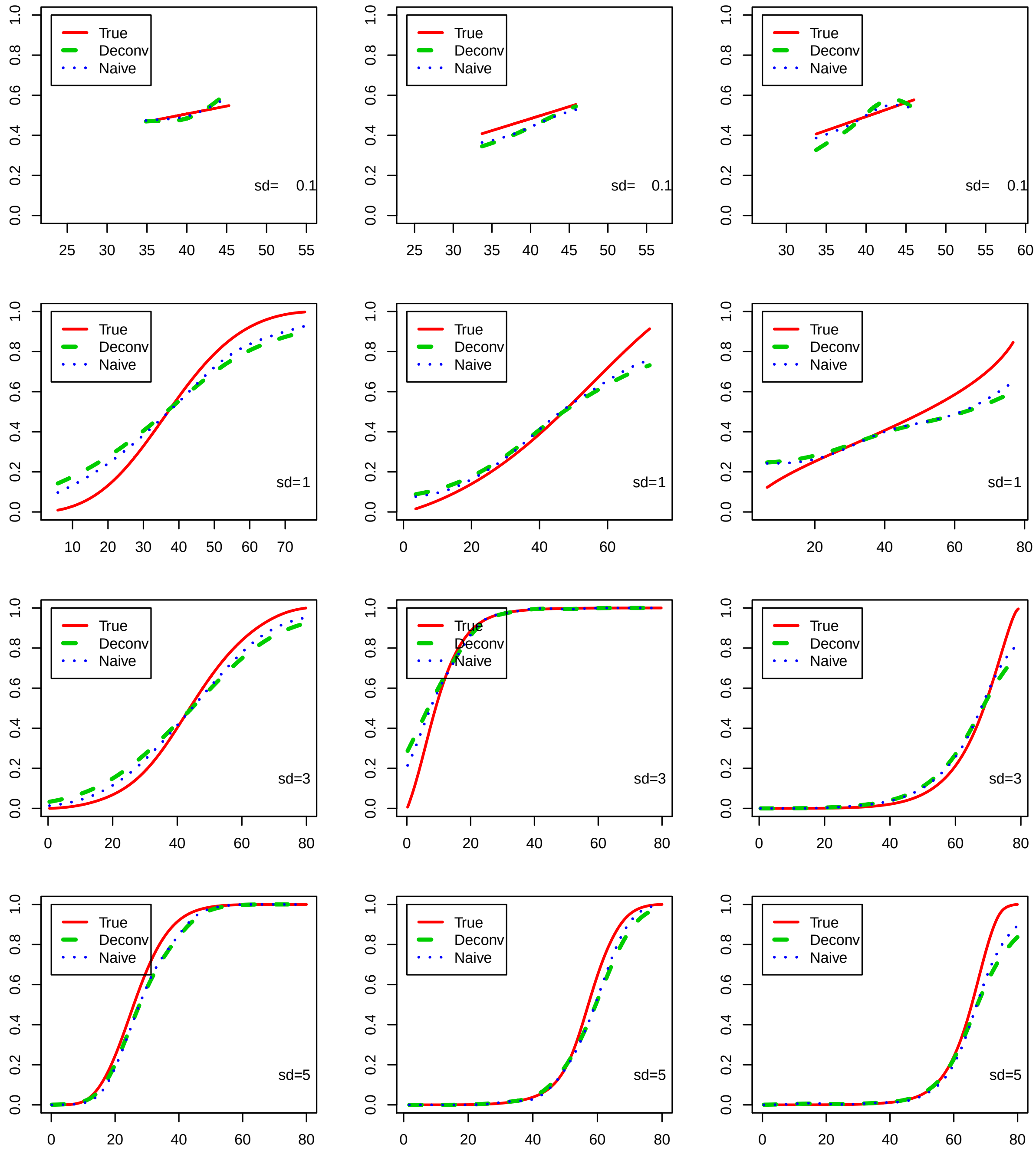

Figure 6. **Item response curve (IRC) plots of several items in a test of length 80 and a sample size of 1,000. In each plot, the x-axis indicates the true score, and the y-axis is the probability of answering the item right for a true score of x. The ability standard deviations varies among 0.1, 1, 3, and 5, as depicted at the bottom right corner of each plot.**

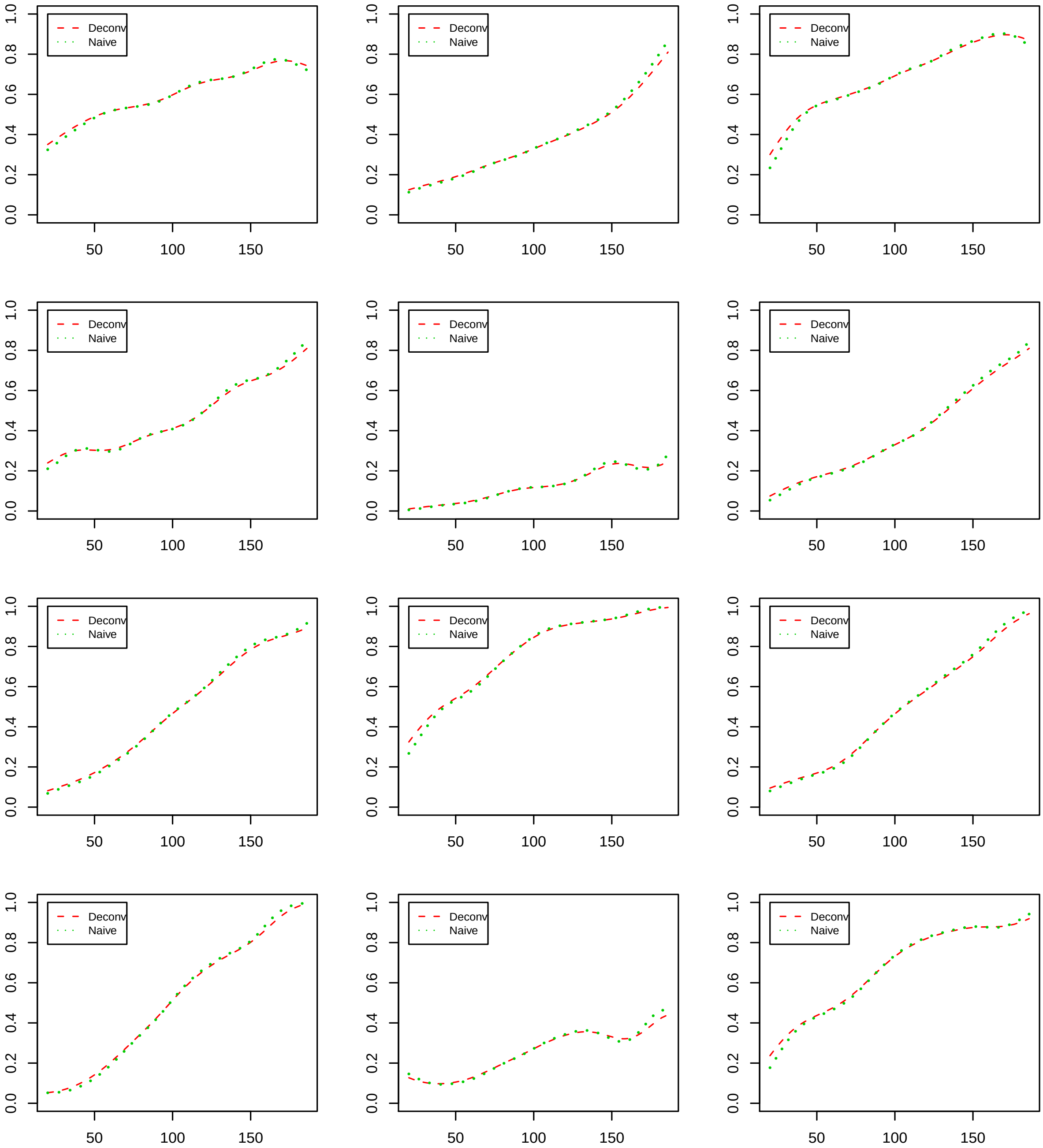

***Figure* 7. Item response curve (IRC) estimations for some items in a real test. The standard error of measurement for Ramsay's estimator (SEM.r) = 6.9. In each plot, the x-axis indicates the true score, and the y-axis is the probability of answering the item right for a score of x.**

naive kernel estimation. Such comparisons (results not shown) also showed virtually no difference between the two methods.

4 Discussion

We investigated the deconvolution method under a variety of conditions to correct the influence of measurement error. The naive method and the deconvolution method produce similar results in our study. The similarity may be attributed to the relatively small SEM ($\sigma_\epsilon \leq 10$) and the relatively large bandwidth ($h > 3$). In this case, the naive kernel function $K(\cdot)$ and the modified kernel function $K^*(\cdot)$ in (3) are close, and thus the two methods yield similar estimations.

When a study's main focus is to investigate an item's IRC property (i.e., whether an item possesses the property that the test takers' chance of obtaining the right answer increases with his or her ability), the naive kernel estimation is competitive compared to other statistical methods with error correction. The deconvolution method provides an asymptotic unbiased estimation; however, the difficulty lies in the unknown distributions of measurement error and unavailable optimal bandwidth choices in practice.

Assuming that the variance of measurement error is a constant is another limitation of the deconvolution method. The measurement error has a heterogenous distribution in many item response IRT models. It is worth investigating whether significant improvement can be expected by using regression models with correction of heterogenous measurement error.

References

Carroll, R.J., Maca, J.D., & Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika, 86*, 541–554.

Carroll, R., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). London, England: Chapman & Hall.

Delaigle, A., Fan, J., & Carroll, R. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association, 104*, 348–359.

Douglas, J., & Cohen, A. S. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25*, 234–243.

Durrett, R. (1995). *Probability: Theory and examples* (2nd ed.). New York, NY: Duxbury Press.

Fan, J., & Troung, Y. (1993). Nonparametric regression with errors in variables. *Annals of Statistics, 21*, 1900–1925.

Haertel, E. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Kolen, M., Zeng, L., & Hanson, B. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*, 129–140.

Lee, Y. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*, 121–134.

Livingston, S., & Dorans, N. (2004). *A graphical approach to item analysis* (ETS Research Report No. RR-04-10). Princeton, NJ: ETS.

Meijer, R. (2004). *Investigating the quality of items in CAT using nonparametric IRT* (LSAC Computerized Testing Report 04-05). Newtwon, PA: Law School Admission Council, Inc.

Ramsay, J. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.

Ramsay, J. (1998). *TestGraf: a program for the graphical analysis of multiple choice test and questionare data.* Retrieved from ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/TestGraf98.pdf.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, *90*, 1257–1270.

Sijtsma, K. (1998). Methodology review: nonparametric IRT approaches to the analysis of dichotomours item scores. *Applied Psychological Measurement, 22*, 3–31.

Silverman, B. (1986). *Density estimation for statistics and data analysis.* London, England: Chapman & Hall.

Stein, E., & Weiss, G. (1971). *Introduction to Fourier analysis on Euclidean spaces.* Princeton, NJ: Princeton University Press.

Wand, M. (1998). Finite sample performance of deconvolving density estimators. *Statistics & Probability letters, 37*, 131–139.

Wand, M., & Jones, M. (1995). *Kernel smoothing.* London, England: Champman & Hall.

Appendix

Measurement Error

To apply the deconvolution method, one needs to know the distribution of the measurement error. Here, we discuss asymptotic results of error distribution in both CTT and IRT models.

Measurement Error in Classical Test Theory (CTT)

Let n be the number of examinees, and J be the number of items on the test. In CTT, the observed score Z and the true score X have the following relationship:

$$Z = X + \epsilon$$

where ϵ is the measurement error independent of X.

Assumption 1. $Z = \sum_{j=1}^{J} Z_j$, and $X = \sum_{j=1}^{J} X_j$, where Z_j and X_j are the item score and true item score, respectively.

Assumption 2. $Z_j - X_j$ are independent of $Z_i - X_i$ for $j \neq i$.

Notice that $\{Z_j - X_j, j = 1, \cdots, J\}$ are bounded and independent random variables by Assumptions 1 and 2. Then, by the Lindeberg-Feller theorem (Durrett, 1995), for $J \to \infty$ (that is, for a very long test),

$$\frac{Z - X}{\sigma_\epsilon} = \frac{\sum_{j=1}^{J}(Z_j - X_j)}{\sigma_\epsilon} \Longrightarrow N(0, 1). \tag{A1}$$

Assumptions 1 and 2 are reasonable conditions. Assumption 1 says that the test score is a sum of item scores while Assumption 2 says that the measurement errors are independent of each other.

Since X is not observable, σ_ϵ can not be calculated directly. However, the test reliability γ can be estimated in many ways (Haertel, 2006), and then σ_ϵ can be estimated as

$$\hat{\sigma}_\epsilon = \sigma_Z \sqrt{1 - \hat{\gamma}}. \tag{A2}$$

Measurement Error in Item Response Theory (IRT)

An important difference between CTT and IRT is the treatment of measurement error. CTT assumes that the variance of error is the same for each examinee, but IRT allows it to vary. The unidimensionality, local independence, and monotonic increment of the item response function (IRF) are assumed here conventionally. The observed score Z and the true score X have the following relationship for given ability θ:

$$Z(\theta) = \sum_{j=1}^{J} Z_j(\theta), \qquad X(\theta) = E(Z(\theta)) = \sum_{j=1}^{J} p_j(\theta),$$

where $Z_j(\theta)$ is the dichotomous item score of the examinee on item j, $E(\cdot)$ is the expectation operator, and $p_j(\theta)$ is the probability of obtaining the right answer on item j for an examinee with ability θ. Let $G(\theta)$ be the distribution function of θ with mean μ_θ and standard deviation σ_θ. Denote the conditional measurement error variance given θ as

$$\sigma_\epsilon^2(\theta) = E[(Z-X)^2|\theta] = E[(Z - E(Z|\theta)|\theta]^2 = \mathrm{Var}(Z|\theta)$$

and the unconditional variance of error can be expressed as (Kolen, Zeng, & Hanson, 1996)

$$\sigma_\epsilon^2 = \int_\theta \sigma^2(Z|\theta)dG(\theta) = \sum_{j=1}^{J} \int_\theta p_j(\theta)\big(1 - p_j(\theta)\big)dG(\theta). \tag{A3}$$

For a long test with fixed item parameters, that is, when $J \to \infty$, the standardized score given θ is given by

$$\epsilon(\theta) := \frac{Z(\theta) - X(\theta)}{\sigma_\epsilon(\theta)} = \frac{1}{\sigma_\epsilon(\theta)} \sum_{j=1}^{J} (Z_j(\theta) - p_j(\theta)) \Longrightarrow N(0,1), \tag{A4}$$

by the Lindeberg-Feller theorem again. Note that the variance of error is a function of θ, which is different from (A1) in the CTT model. Also note that the right hand side of (A4) is a random variable independent of θ.

Assumption 3. Suppose that

$$\frac{\mathrm{Var}(\sigma_\epsilon(\theta))}{\sigma_\epsilon^2} \to 0.$$

Then, under Assumption 3, (A4) can be rewritten as

$$\frac{Z(\theta) - X(\theta)}{\sigma_\epsilon} \Longrightarrow N(0,1). \tag{A5}$$

Assumption 3 requires that the ratio of variation of $\sigma_\epsilon^2(\theta)$ is very small as the ability variable θ varies. Under the CTT model, as $\sigma_\epsilon^2(\theta)$ is a constant for all examinees, $\mathrm{Var}(\sigma_\epsilon(\theta)) = 0$; hence this assumption is true. When the ability of the examinees is not too heterogeneous, that is, when $\mathrm{Var}(\sigma_\epsilon(\theta))$ is very small compared to σ_ϵ^2, Assumption 3 is likely to be true too.

Numerical Results

We use the 2PL IRT model to simulate data to investigate the error distributions and SEM. The 2PL model assumes that

$$P(Z_j = 1|\theta) = \frac{e^{\alpha_j \theta - \beta_j}}{1 + e^{\alpha_j \theta - \beta_j}}. \tag{A6}$$

In the simulation, test lengths are 20, 40, and 80 for short, medium and long tests. Sample sizes are 100, 500, 1,000, and 5,000 for small, medium, large, and very large samples. The distributions of ability, β parameter, and α parameter follow $N(0, \sigma_\theta)$, $N(0, \sigma_\theta)$, and $N(1, 0.25)$, respectively. For 2PL IRT models, (A3) becomes

$$\sigma_\epsilon^2 = \sum_{j=1}^{J} \int_\theta \frac{e^{\alpha_j \theta - \beta_j}}{(1 + e^{\alpha_j \theta - \beta_j})^2} \psi(\theta) d\theta, \tag{A7}$$

where $\psi(\theta)$ is the normal density function.

Three SEMs are calculated for each simulated data set: SEM.th is from the theoretical formula (A7) by using Gaussian quadrature approximation of integration; SEM.r is obtained from (A2), where the reliability is estimated using the Cronbach's alpha method (e.g., Haertel, 2006); and SEM.em is the empirical SEM using data and true scores, that is, SEM.em $= \sum_{i=1}^{n} (Z_i - X_i)^2/(n-1)$, where Z_i and X_i are the observed score and true score for examinee i, and n is the number of examinees. The three SEMs are compared in Table A1. None of the three SEMs are affected by the sample size, but all are affected by the test length and the variability of the ability. As expected, as tests become longer, SEM increases; when the sample ability is more heterogeneous, SEM is slightly smaller (because the test reliability is larger for a more heterogeneous population). Overall, the SEMs are relatively comparable across the three different methods of calculation.

The variance and mean square ratios of $\sigma_\epsilon(\theta)$, as in the Assumption 3, are displayed in Table A2. The ratio is positively proportional to the size of ability variation. Verification of normality of the error distribution is displayed in QQ plots in Figures A1 to A3. The points in the Q Q plot are formed by pairs of estimated quantiles from the data $(\epsilon_i, \cdots, \epsilon_n)$ and estimated quantiles from n observations of a normal distribution $N(0, \sigma_\epsilon)$. Both axes are in units of their respective data sets. If the two sets come from a population with the same distribution, the points should fall approximately along a 45-degree reference line. From these QQ plots, we observe that the

Table A1

Comparison of the Standard Error of Measurement (SEM) for Tests of Length 20, 40, and 80

		SEM.r			SEM.th			SEM.em		
Test length	n\ σ_θ	.1	1	5	.1	1	5	.1	1	5
20	100	2.21	1.92	1.38	2.23	1.92	1.02	2.36	1.97	1.17
	500	2.24	1.94	1.39	2.23	1.97	1.08	2.22	1.86	1.16
	1,000	2.23	1.92	1.40	2.23	1.92	1.12	2.14	1.99	1.05
	5,000	2.23	1.85	1.38	2.23	1.91	1.06	2.26	1.94	1.09
40	100	3.16	2.67	1.95	3.15	2.70	1.43	2.73	2.71	1.42
	500	3.16	2.65	1.98	3.15	2.70	1.51	3.23	2.76	1.49
	1,000	3.15	2.75	1.97	3.16	2.67	1.47	3.22	2.63	1.45
	5,000	3.15	2.73	1.96	3.15	2.71	1.38	3.11	2.63	1.56
80	100	4.46	3.87	2.77	4.46	3.81	2.17	4.45	3.83	2.17
	500	4.46	3.85	2.77	4.46	3.81	2.13	4.53	3.44	2.19
	1,000	4.46	3.80	2.77	4.46	3.80	2.01	4.57	3.94	2.38
	5,000	4.46	3.82	2.75	4.46	3.90	2.03	4.51	3.88	2.12

Note. SEM.em = empirical SEM; SEM.r = SEM obtained from (A2), where the reliability is estimated using the Cronbach's alpha method; SEM.th = SEM obtained from theoretical formula (A7) by using Gaussian quadrature approximation of integration.

empirical distribution of error does not deviate much from a normal distribution under a variety of conditions when samples are relatively large ($n \geq 500$).

Table A2

Ratio of Variance and Mean-Square for $\sigma_\epsilon(\theta)$

	J = 20			J = 40			J = 80		
n\ σ_θ	.1	1	5	.1	1	5	.1	1	5
100	.0023	.1380	.2701	.0023	.1034	.2384	.0028	.1642	.3560
500	.0026	.1692	.3958	.0026	.1237	.2522	.0024	.1163	.3202
1,000	.0026	.1913	.4312	.0027	.1531	.3144	.0025	.1235	.3318
5,000	.0026	.1226	.2620	.0024	.1348	.2388	.0024	.1184	.3195

Note. J = test length.

From IRT simulations, the measurement error can be approximated by a normal distribution for a moderate long test ($J \geq 40$) and a medium sized population ($n \geq 500$). Even for shorter tests with smaller sample sizes, the normal approximation is still acceptable sometimes. However, whether real data have such a property is unknown since the true scores are unobservable.

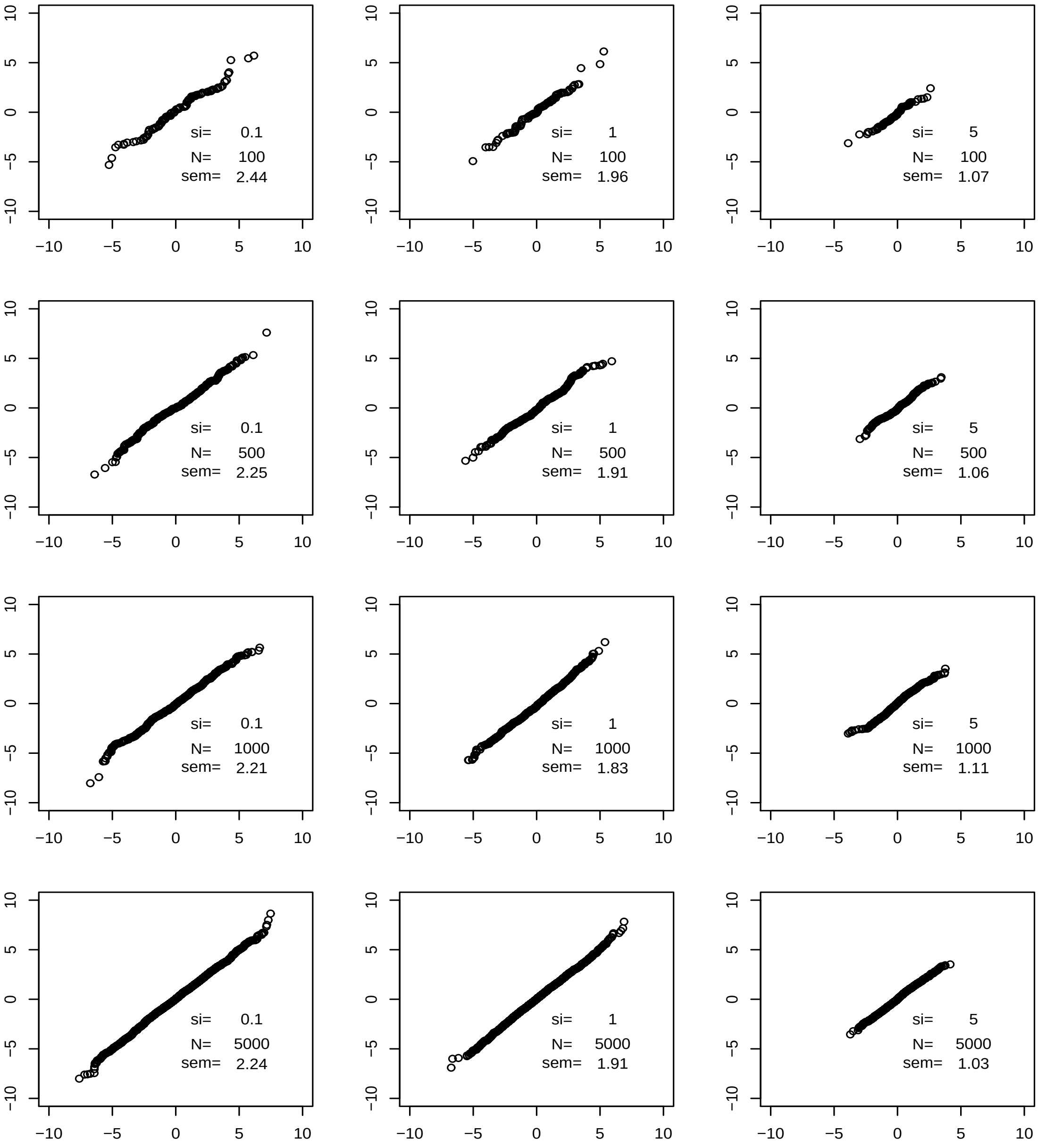

***Figure* A1. Error distribution comparison with a normal distribution for a test of length 80. In each plot, si is the standard deviation of the ability, *N* is the sample size in the simulation, and SEM.em is the empirical SEM.**

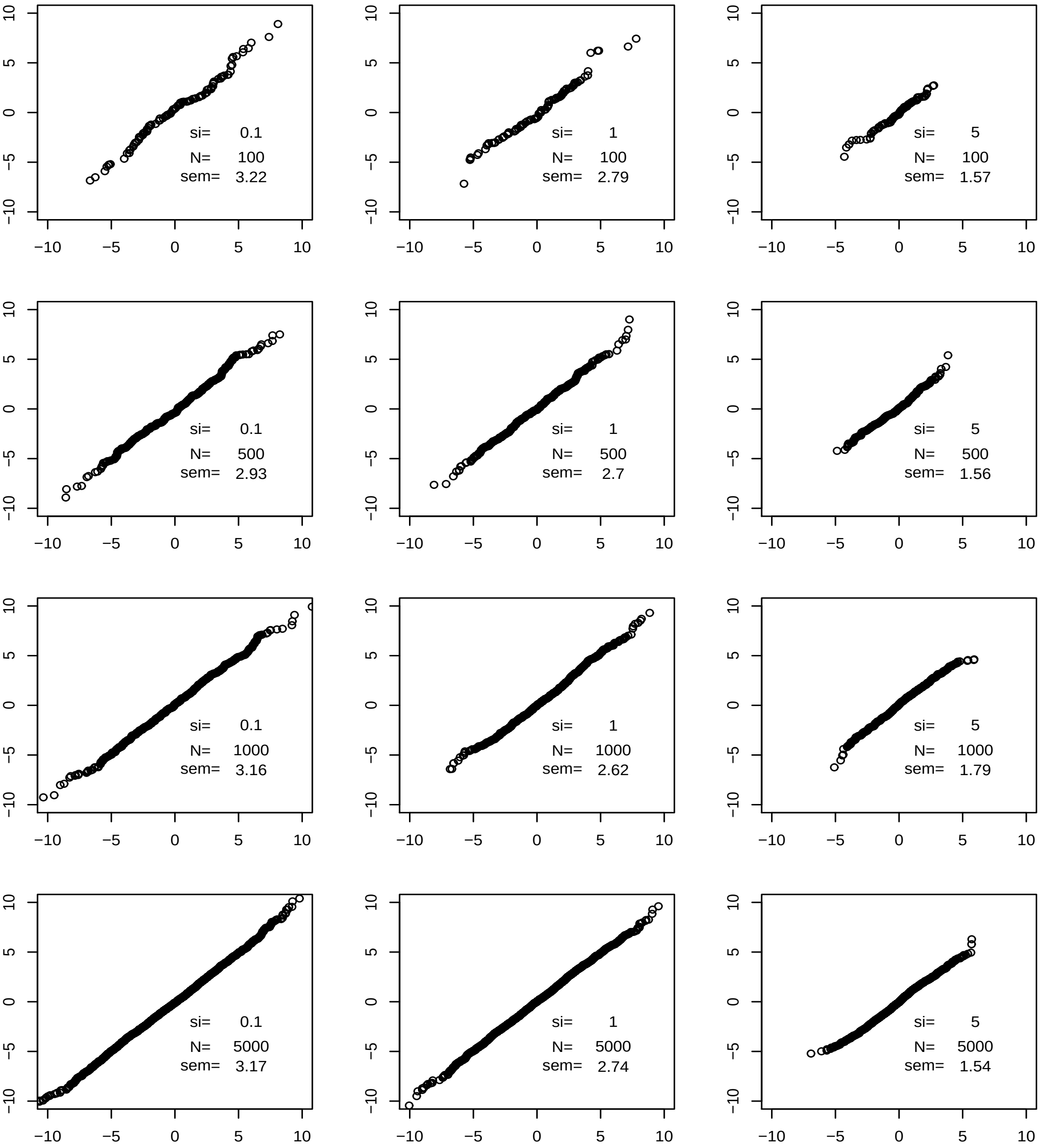

***Figure* A2. Error distribution comparison with a normal distribution for a test of length 80. In each plot, si is the standard deviation of the ability, *N* is the sample size in the simulation, and SEM.em is the empirical SEM.**

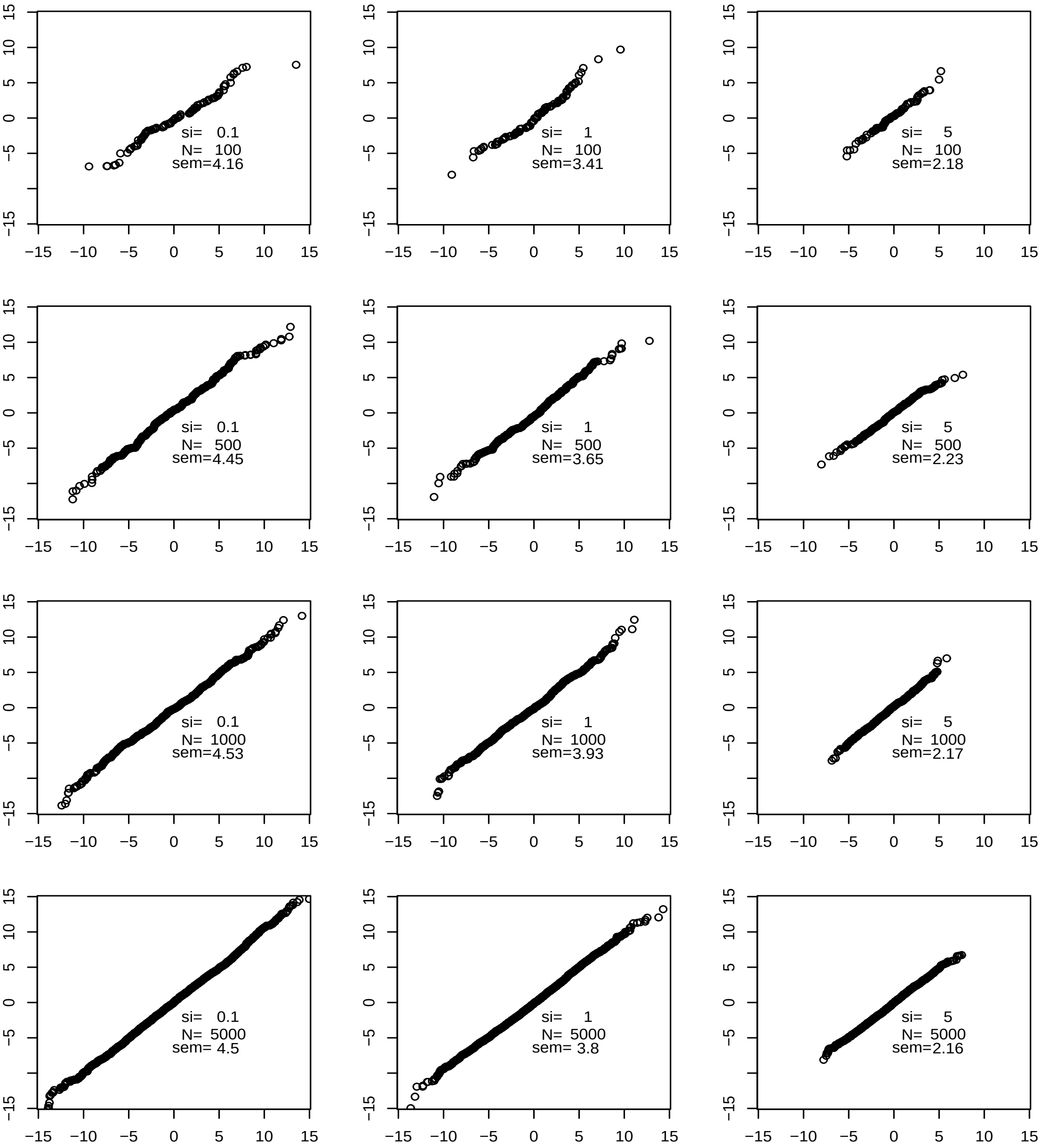

***Figure* A3. Error distribution comparison with a normal distribution for a test of length 80. In each plot, si is the standard deviation of the ability, *N* is the sample size in the simulation, and SEM.em is the empirical SEM.**