Effect of Scale Adjustment on the Comparison of Item and Ability Parameters

Michelle Liou
Academia Sinica

The standardized mean-squared difference (SMSD) has been used for summarizing the bias of parameter estimates in the three-parameter logistic (3PL) model. Due to the indeterminacy problem of the 3PL model, researchers must select a common scale for comparing the theoretical and estimated parameters. The use of different scales can yield noncomparable SMSD values, which in turn can affect the comparison of bias between different parameters. This research used three methods for selecting the common scale. Through a simulation, the three scaling methods were used to numerically demonstrate their effect on SMSD values. *Index terms: equating, indeterminacy problem, Samejima scale, standardized mean-squared difference, Stocking and Lord scale, three-parameter logistic model.*

In the three-parameter logistic (3PL) model, the probability of a person with ability θ answering an item correctly is defined as

$$P_i = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]} \quad , \tag{1}$$

where a_i, b_i, and c_i are the item discrimination, difficulty, and guessing parameters, respectively, for the ith item. In the model, the determination of a scale for the a, b, and θ parameters is purely arbitrary. If the parameters are linearly transformed with the slope α and intercept β, such that $a_i^* = a_i/\alpha$, $b_i^* = \alpha b_i + \beta$ and $\theta^* = \alpha\theta + \beta$, the value of P_i is unchanged. This has been termed the indeterminacy problem associated with the 3PL model (Lord, 1980). Because the identification of the parameter scale is arbitrary, parameters calibrated with different datasets are not directly comparable. In order to make them comparable, some appropriate scale adjustment or equating is needed to place all of the estimated item response functions (IRF) on a common scale.

The question of statistical accuracy of parameter estimates in the 3PL model is of practical and theoretical concern. There are many instances in which this question has been answered through evaluation of the standardized mean squared difference (SMSD) for the a, b, c and θ parameters separately (e.g., Drasgow & Parsons, 1983; Yen, 1984). If there are m parameter values under consideration, the SMSD value (Yen, 1984) is defined as the ratio between the two quantities, $\Sigma_i^m(\omega_i - \hat{\omega}_i)^2/m$ and $(\sigma_\omega^2 + \sigma_{\hat{\omega}}^2)/2$, where ω denotes the theoretical parameter, $\hat{\omega}$ the parameter estimate, and σ^2 the variance of the m parameter values. Due to the indeterminacy problem of the 3PL model, researchers must select a common scale for comparing the theoretical parameters and corresponding estimates. The estimated bias of individual parameters could be confounded with the common scale selected, however, because the absolute and relative sizes of the SMSD are sensitive to the predetermined scale.

This research considered three scaling procedures that have been widely used by researchers as methods for finding the common scale: the standardized scale (i.e., the mean and standard deviation of the theoretical and estimated θ are 0 and 1, respectively), the Stocking and Lord (1983) scale, and

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 3, September 1990, pp. 313-321
© Copyright 1990 Applied Psychological Measurement Inc.
0146-6216/90/030313-09$1.90

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

the Samejima (1984) scale. Two features that differ in these three scales are (1) the definition of the criterion function in deciding the scaling constants, α and β (i.e., the least squares or the principal axis criterion), and (2) the source of calibration data used in the criterion function (i.e., item or ability estimates). Within the same dataset these two features would be likely to interact, and the three scales could result in different patterns of the SMSD values between individual parameters.

This research was designed to investigate the effect of scale selection on error in calibrating item and ability parameters, with particular reference to the SMSD statistic; it was not intended to provide a definitive answer to the substantive question regarding the relative merits of the scaling procedures. Therefore, the three scales used serve only as vehicles for finding a common scale and for demonstrating the problem of scale selection.

Method and Results

Scale Selection

In order to illustrate the problem of scale selection, a simulation study was conducted by assuming the 3PL model. Three hypothetical tests were simulated, each containing $m = 20$, 40, and 60 five-choice items. The 20- and 40-item tests were comprised of items 1–20 and 1–40, respectively, from the 60-item test. The a, b, and c parameters for the 60 items were generated randomly from the ranges and quartiles as specified by Lord (1983, p. 238). The simulated item parameters are listed in Table 1.

Five hundred random deviates ($n = 500$) were generated from a standard normal distribution using the IMSL subroutine GGNPM (International Mathematical and Statistical Libraries, 1982). These random numbers were assumed to be the true θ values. For the ith item and the jth θ, the P_{ij} value was computed and was then used to generate binary item scores. An item score μ_{ij} was simulated by generating first a random number from a uniform distribution using the IMSL subroutine GGUBS, and then assigning a score of 1 to μ_{ij} if the random number was less than or equal to P_{ij}, or a score of 0 otherwise. All the simulated item scores on the three tests were calibrated by the LOGIST-5 computer program (Wingersky, Barton, & Lord, 1982) using default options.

The simulation of item scores and data calibration were replicated five times with the same true item parameters and θ values. The theoretical θ values and the corresponding $\hat{\theta}$ values from the first simulation replication are plotted in Figure 1. The average test mean, test variance, and the internal consistency index (i.e., Cronbach's alpha) computed from number-correct scores over the five replications are listed in Table 2.

The index $\pi_{ij} = a_i(\theta_j - b_i)$ was defined for $i = 1, \ldots, m$ and $j = 1, \ldots, 500$. Because the π and c parameters are unaffected by linear transformation of the item and θ parameters, the SMSD values for these two parameters were computed directly. Table 2 also contains the average SMSD values for π and c over the five replications. The simulation results suggest that an increase in test length decreases the SMSD value for the π index, but not for the guessing parameter; collectively, the π index was much more accurately estimated than the guessing parameter.

The Standardized Scale

Method. Researchers commonly assume that the mean and SD of the distribution of θ (μ_θ and s_θ) can be approximated by those of $\hat{\theta}$ (Drasgow & Parsons, 1983; Samejima, 1986). Following this assumption, the origin and unit of $\hat{\theta}$ can be set at the sample mean $\mu_{\hat{\theta}}$ and SD $s_{\hat{\theta}}$ so that the standardized $\hat{\theta}$ values will be comparable to the θ values, whose origin and unit have also been set at μ_θ and s_θ, respectively. The a and b parameters, along with their corresponding estimates, are also adjusted according to the standardized scale.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Theoretical and Estimated Item Parameters for the 60 Simulation Items Based on 20, 40, and 60 Items Used in Data Calibration

		$\hat{a}$				$\hat{b}$				$\hat{c}$		
Item	a	20	40	60	b	20	40	60	c	20	40	60
1	.72	1.04	.96	.87	.53	.26	.39	.37	.14	.08	.13	.11
2	1.41	1.61	1.69	1.34	1.68	1.34	1.38	1.48	.18	.20	.20	.20
3	.78	1.08	.98	1.08	.75	.68	.70	.77	.02	.05	.04	.07
4	1.01	.92	1.05	1.19	1.37	.90	1.03	1.14	.18	.11	.15	.19
5	1.46	1.77	1.52	1.37	1.04	.99	1.01	1.08	.13	.16	.15	.15
6	.50	.63	.42	.44	.15	.63	.21	.35	.05	.26	.11	.15
7	.64	1.33	1.08	1.01	.19	.49	.52	.52	.10	.27	.26	.26
8	.72	.55	.49	.49	-3.73	-4.44	-4.38	-4.21	.03	.15	.11	.15
9	.83	.90	.53	.72	1.36	1.09	.92	1.09	.15	.13	.00	.09
10	.76	.90	.86	. 86	-.15	.02	.04	.02	.02	.10	.10	.09
11	.49	.55	.30	.29	2.23	1.90	1.89	1.98	.14	.13	.00	.00
12	.82	.68	.48	.55	1.91	1.50	1.64	1.78	.18	.09	.04	.09
13	.98	.91	1.24	1.08	.00	-.19	.17	.10	.15	.07	.23	.20
14	1.70	.62	.84	.74	-2.65	-6.53	-4.16	-4.25	.17	.15	.11	.15
15	.49	.81	.60	.57	-.10	.15	-.11	-.02	.11	.22	.11	.15
16	1.01	1.73	1.47	1.13	-.32	-.14	-.18	-.38	.14	.24	.22	.13
17	.51	.76	.70	.63	.23	.05	.19	.30	.06	.07	.11	.15
18	.78	1.54	1.62	1.55	1.73	1.26	1.32	1.33	.33	.34	.35	.35
19	.97	1.06	.90	.92	-.65	-.66	-.76	-.70	.16	.15	.11	.15
20	1.40	1.17	.78	.85	1.27	.81	.82	1.00	.45	.33	.29	.34
21	.82		.61	.57	2.21		2.72	2.85	.15		.18	.18
22	.45		.69	.58	2.31		1.84	1.93	.13		.16	.14
23	1.33		2.00	2.00	.50		.40	.39	.19		.16	.16
24	1.33		1.54	1.67	-1.35		-1.22	-1.09	.04		.11	.21
25	.91		1.17	1.10	.00		.11	.18	.14		.14	.17
26	.60		.29	.28	1.70		.00	.21	.45		.11	.15
27	.55		1.25	.82	1.52		1.39	1.57	.14		.23	.20
28	.94		.77	.74	1.55		1.40	1.40	.20		.14	.13
29	.95		.50	.62	-2.92		-4.79	-3.90	.14		.11	.15
30	1.52		2.00	2.00	1.88		1.70	1.79	.42		.40	.41
31	1.45		.67	.77	-2.57		-4.92	-4.32	.13		.11	.15
32	1.17		1.44	1.16	.45		.37	.36	.15		.16	.15
33	1.10		1.14	1.08	.12		.10	.04	.18		.20	.17
34	.70		.72	.72	-.06		-.18	-.10	.15		.11	.15
35	.88		.34	.34	-3.70		-6.95	-6.83	.42		.11	.15
36	.51		2.00	.45	2.15		2.02	2.57	.19		.32	.23
37	.55		.63	.68	-3.45		-3.37	-3.07	.15		.11	.15
38	.47		.19	.20	.47		-2.14	-1.83	.41		.11	.15
39	1.25		.53	.46	-2.93		-5.53	-6.02	.44		.11	.15
40	.74		.74	.73	.27		.01	.11	.14		.11	.15
41	.75			.52	-3.65			-5.06	.03			.15
42	.79			.60	.52			-.52	.15			.06
43	.82			.97	2.23			1.90	.05			.05
44	1.30			1.46	.34			.40	.07			.09
45	.80			1.04	-1.61			-1.57	.14			.15
46	.60			1.11	.71			1.04	.14			.35
47	.88			1.16	1.27			1.23	.16			.15

continued on the next page

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1, continued
Theoretical and Estimated Item Parameters for the 60 Simulation Items Based on 20, 40, and 60 Items Used in Data Calibration

		$\hat{a}$				$\hat{b}$				$\hat{c}$		
Item	a	20	40	60	b	20	40	60	c	20	40	60
48	1.31			1.08	-1.97			-2.39	.13			.15
49	.90			.63	.36			-.31	.30			.15
50	1.02			1.40	.39			.55	.09			.19
51	1.64			1.46	1.38			1.32	.08			.05
52	1.22			1.52	1.55			1.39	.14			.13
53	.77			.97	1.29			1.21	.10			.14
54	.96			1.22	1.33			1.09	.39			.37
55	.90			1.21	.90			.83	.07			.10
56	.95			.67	1.80			1.52	.19			.08
57	.64			.59	-.23			-.21	.13			.15
58	.78			.73	1.20			1.23	.13			.13
59	.68			1.14	.24			.72	.38			.48
60	.82			.33	2.31			3.65	.14			.09

In calibrating the simulation data for this study, $\hat{\theta}$ values that exceeded 3 in absolute value after the last iteration were excluded during the estimation of μ_θ and s_θ (this is also the default option in LOGIST). Because $\hat{\theta}$ obtained from LOGIST had been standardized internally during data calibration, no further transformation was necessary for all the parameter estimates. The true θ values were standardized to mean = 0 and SD = 1. The slope and intercept used in the standardization were applied to rescaling the a and b parameters. The SMSD value between the rescaled parameters and corresponding estimates from LOGIST were computed for each replication.

Results. The average of the five SMSD values for the three tests are also listed in Table 2. The simulation results from the standardized scale method suggest that an increase in test length significantly improves the accuracy of the b and θ estimates, and slightly improves the accuracy of the a estimates. The SMSD values for the a parameter were generally larger than those of the b and θ parameters. Collectively, the b parameter was most accurately estimated, followed by the θ parameter, and the a parameter was least accurately estimated.

The Stocking and Lord Scale

Method. Stocking and Lord (1983) designed a method for equating two calibrations of the same test, and others (e.g., Yen, 1984) have used it to obtain a common scale for evaluating the SMSD values. The Stocking and Lord scale is intended to incorporate more of the information available from data calibration. For example, one calibration of an item yields an estimated IRF $P_i(\hat{a}_i, \hat{b}_i, \hat{c}_i|\theta)$, which is an approximation to $P_i(a_i, b_i, c_i|\theta)$. If the estimate is error-free, the proper choice of α and β for the linear transformation would cause these two curves to coincide. In practice, α and β should be selected for a suitable group of trait values, θ^0, so that the average squared difference between the true IRFs and their estimates is as small as possible. The criterion function to be minimized in deciding the scale is:

$$F(\alpha, \beta) = \frac{1}{m}\sum_{j}^{n}\left\{\sum_{i}^{m} P_{ij}(a_i, b_i, c_i|\theta_j^0) - \sum_{i}^{m} P_{ij}(a_i^*, \hat{b}_i^*, \hat{c}_i|\theta_j^0)\right\}^2 , \quad (2)$$

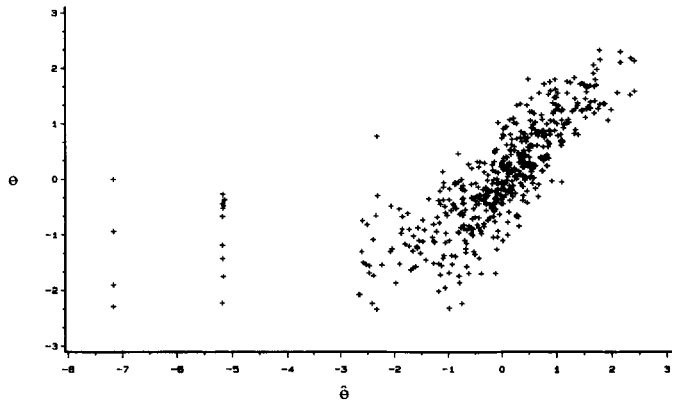
where

$$\hat{a}_i^* = \frac{1}{\alpha}\hat{a}_i \text{ and } \hat{b}_i^* = \alpha\hat{b}_i + \beta \quad . \quad (3)$$

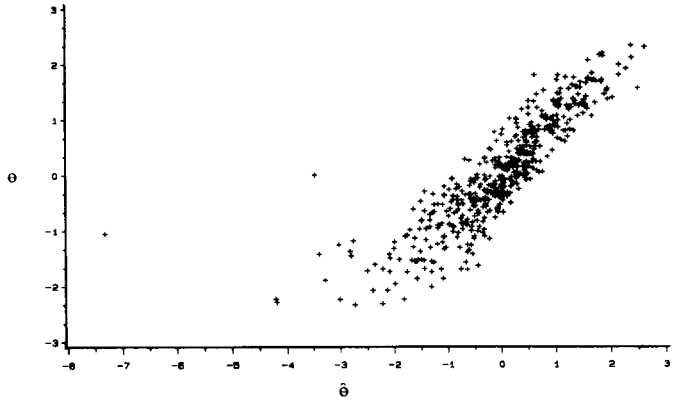
Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Figure 1
Theoretical Versus Estimated θ Parameters

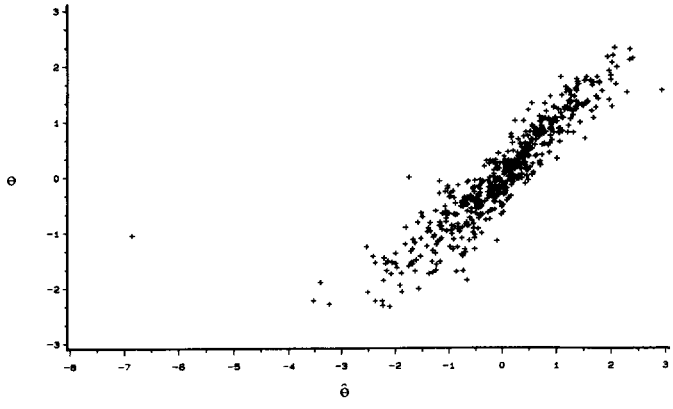
(a) 20-Item Test



(b) 40-Item Test



(c) 60-Item Test



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
SMSD Values, Average Number-Correct Score Mean and Standard Deviation, and Cronbach's Alpha Over Five Simulation Replications for 30-, 40-, and 60-Item Tests

Method and Parameter	Number of Items		
	20	40	60
π	.449	.271	.182
c	.730	1.097	.931
Standardized Scale			
a	1.301	1.089	.938
b	.216	.158	.096
θ	.615	.341	.267
Stocking and Lord Scale			
a	.843	.872	.825
b	.358	.226	.124
θ	.730	.365	.276
Samejima Scale			
a	.923	.976	.851
b	.295	.177	.105
θ	.684	.341	.266
Mean	10.146	22.255	31.564
Variance	13.038	36.258	81.006
Alpha	.718	.811	.874

The minimization of this function normally takes an iterative computer search.

The 500 trait values, θ^0, needed to determine the transformation constants, α and β, were generated by GGNPM with mean = 0 and SD = 1; these trait values were generated independently of the true θ values that were used in simulating the item responses. Items with difficulty estimates outside the range of -5.0,5.0 were excluded in calculating the common scale, because the excluded estimates contained a relatively large standard error of estimation. For the remaining items, the theoretical and estimated item parameters and the 500 trait values were substituted into the criterion function.

The Gauss-Newton procedure (Daniels, 1978; Scales, 1985) was used to minimize $F(\alpha,\beta)$. After α and β were found, the estimated parameters, $\hat{a}$, $\hat{b}$, and $\hat{\theta}$, were all adjusted by the slope and intercept. The minimization process was repeated for the simulation data from the five replications. In each replication, the SMSD values between the true parameters and the rescaled estimates were computed.

Results. The average SMSD values for each parameter are listed in Table 2. The simulation results indicate that the accuracy of the b and θ estimates is greatly improved by increasing the test length from 20 to 40 items. However, the accuracy of the a estimate was improved only when the test was increased to 60 items. In an IRF, the a parameter is strongly related to the c parameter. Therefore, the bias pattern between two tests of different lengths is more likely to be the same for both $\hat{a}$ and $\hat{c}$, and simulation results from the Stocking and Lord scale support this rationale. In other words, the accuracy of both $\hat{a}$ and $\hat{c}$ was not improved when the test was lengthened to 40 items, and it was improved when the test was lengthened to 60 items.

The difference between the SMSD values for the a and θ parameters in the Stocking and Lord scale was not as large as the results from the standardized scale, especially for the 20-item test. The former scale also resulted in larger SMSD values for the b and θ parameters than did the standardized scale.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

It was possible that simulation results from the Stocking and Lord scale could change if the number of θ^0 values was increased from 500 to 1,000. Therefore, a and b were recalculated with 1,000 θ^0 values for the five replications. The simulation results from the replicated study differed in no important ways from the results listed in Table 2.

The Samejima Scale

Method. Samejima (1984) proposed an iterative method that uses the first principal component as the best fitted linear relationship for the scatter diagram of either $\hat{a}$ or $\hat{b}$ from two calibrations of the same test. The method also can be used to find a common scale for evaluating the SMSD. For example, the a parameter is proportional to the slope of the IRF at $\theta = b$. If $\hat{a}$ is error-free, the second principal component of $\hat{a}$ and a should degenerate, and $\hat{a}_i = (s_{\hat{\theta}}/s_{\theta})a_i$ $(i = 1, \ldots, m)$ with $s_{\hat{\theta}}$ and s_{θ} as the standard deviations of the 500 $\hat{\theta}$ and θ values, respectively. Samejima's proposal was used in this study because the a estimate contained error, and the first principal component worked as the best fitted linear relationship between a and $\hat{a}$, or equivalently, as an estimate of $s_{\theta}/s_{\hat{\theta}}$.

In order to find the principal component, it was necessary to minimize the criterion function

$$F'[\cos(\varphi), \sin(\varphi)] = \sum_{i}^{m}[\cos(\varphi)a_i - \sin(\varphi)\hat{a}]^2 \quad . \tag{4}$$

The rotation of the axis locating the a estimates was through an angle φ. The minimum of $F'[\cos(\varphi), \sin(\varphi)]$ was found by evaluating

$$\frac{S_{\theta}}{S_{\hat{\theta}}} \approx \tan(\varphi) = \frac{\left(\sum a_i^2 - \sum \hat{a}_i^2\right) \pm \left\{\left(\sum a_i^2 - \sum \hat{a}_i^2\right)^2 + 4\left(\sum a_i \hat{a}_i\right)^2\right\}^{1/2}}{2\sum a_i \hat{a}_i} \quad , \tag{5}$$

where $\tan(\varphi)$ was chosen to be positive. Then the $\cos(\varphi)$ and $\sin(\varphi)$ values were determined by computing $\cos\{\text{arc}[\tan(\varphi)]\}$ and $\sin\{\text{arc}[\tan(\varphi)]\}$, respectively.

In the above method, equal weights are placed on the values of a and $\hat{a}$, which are normally determined on different ability scales. If a unit in the scale for ability estimates is increased, the values of $\hat{a}$ and their errors also would become larger. In calculating $\tan(\varphi)$ in Equation 5, the value of $\hat{a}$ that is based on the larger scale unit is more likely to be penalized (Samejima, 1984). Therefore, a weighting factor was needed to adjust the scale difference. If the true ratio $s_{\theta}/s_{\hat{\theta}}$, or its equivalent, $\tan(\varphi)$, is known, the weighted parameters $\hat{a}^* = \sin(\varphi)\hat{a}$ and $a^* = \cos(\varphi)a$ could be substituted into Equation 5 to obtain $\tan(\varphi)$. Unlike a and $\hat{a}$, a^* and $\hat{a}^*$ are comparable.

In practice, however, the value of $s_{\theta}/s_{\hat{\theta}}$ is unknown. An estimated ratio based on original a and $\hat{a}$ can be used as its initial estimate. Samejima suggests the following iterative procedure for estimating $s_{\theta}/s_{\hat{\theta}}$:

Step 1. Set $j = 1$, $\cos(\varphi) = 1.0$, $\sin(\varphi) = 1.0$, and $\tan(\varphi) = 1.0$.

Step 2. Set $a_i = \cos(\varphi) \times a_i$, and $\hat{a}_i = \sin(\varphi) \times \hat{a}_i$, $(i = 1, \ldots, m)$.

Step 3. Compute $\tan(\varphi)^{(j)}$, $\cos(\varphi)$ and $\sin(\varphi)$, where j denotes the iteration number.

Step 4. Set $\tan(\varphi) = \tan(\varphi)^{(j)} \times \tan(\varphi)$.

Step 5. If $\tan(\varphi)^{(j)} \neq 1.0$, $j = j+1$; go to Step 2.

The iterative procedure terminates at $\tan(\varphi)^{(j)} = 1$. The last value of $\tan(\varphi)$ is an estimate of $s_{\theta}/s_{\hat{\theta}}$. Samejima (1984) gives a detailed description of the iterative algorithm.

The b and $\hat{b}$ values can also be used to obtain an estimate for $s_{\theta}/s_{\hat{\theta}}$. For example, if $\hat{b}$ is error-

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

free, then $\hat{b}_i = (s_\theta/s_{\hat{\theta}})b_i + (\mu_\theta - \mu_{\hat{\theta}})/s_{\hat{\theta}}$.

Samejima suggested that

$$F''[\cos(\varphi), \sin(\varphi)] = \sum_i^m \{\sin(\varphi)(b_i - \mu_b) - \cos(\varphi)(\hat{b}_i - \mu_{\hat{b}})\}^2 \quad , \tag{6}$$

be minimized to approximate the ratio $s_\theta/s_{\hat{\theta}}$, where μ_b and $\mu_{\hat{b}}$ are the average of the b and $\hat{b}$ values, respectively. The minimization of Equation 6 is equivalent to estimating $(\mu_\theta - \mu_{\hat{\theta}})/s_{\hat{\theta}}$ with $[\sin(\varphi)\mu_b - \cos(\varphi)\mu_{\hat{b}})]/\cos(\varphi)$. The directional cosine in Equation 6 is found by evaluating

$$\frac{S_\theta}{S_{\hat{\theta}}} \approx \tan(\varphi) = \frac{(\sigma_{\hat{b}}^2 - \sigma_b^2) \pm \{(\sigma_{\hat{b}}^2 - \sigma_b^2)^2 + 4\sigma_{b\hat{b}}^2\}^{1/2}}{2\sigma_{b\hat{b}}} \quad , \tag{7}$$

where σ_b^2 and $\sigma_{\hat{b}}^2$ are the variances of the b and $\hat{b}$ values, respectively, and $\sigma_{b\hat{b}}$ is their covariance. The value of $\tan(\varphi)$ is also chosen to be positive. The aforementioned iterative procedure then was used to find an estimate for $s_\theta/s_{\hat{\theta}}$.

Based on the obtained a and b parameters, two separate estimates of $\tan(\varphi)$ are obtained. Samejima used the geometric mean of the two estimates as the ultimate estimate of the ratio, $s_\theta/s_{\hat{\theta}}$.

For the simulation data, items with difficulty estimates outside the range of -5.0,5.0 were excluded in the process of fitting the linear relationship. The two estimates of $s_\theta/s_{\hat{\theta}}$ were calculated and their geometric mean was found. The theoretical a and b values, accompanied by their estimated values, were rescaled according to the geometric mean. Thus the a and $\hat{a}$ values were rescaled by the slopes, $\cos(\varphi)$ and $\sin(\varphi)$, respectively. Because the b and θ parameters had to be transformed by the same slope and intercept, the slope $\sin(\varphi)$ and intercept $-\sin(\varphi)\mu_b$ were used to adjust both the b and θ values, and $\cos(\varphi)$ and $-\cos(\varphi)\mu_{\hat{b}}$ to adjust the $\hat{b}$ and $\hat{\theta}$ values.

Results. The average SMSD values over the five replications are listed in Table 2. Simulation results from the Samejima scale also suggest that the accuracy of $\hat{a}$ is improved when the test is lengthened to 60 items. However, the absolute SMSD values for the a parameter are slightly larger than those derived from the Stocking and Lord scale. Conversely, the SMSD values for the b and θ parameters have comparable sizes to those derived from the standardized scale.

Discussion

The greatest variability of SMSD values occurred between different scales when the test contained 20 or 40 items. Therefore, scale selection is critical for evaluating the SMSD values when error in calibrating item and θ parameters is large. The three scales—the standardized scale, the Stocking and Lord scale, and the Samejima scale—yielded quite consistent SMSD values for the θ parameter. This suggests that the problem of scale selection is of particular importance when accuracy in calibrating item parameters is of major concern.

The simulation results also suggest that the source of calibration data used for determining a common scale is also important in determining the magnitude of SMSD values. For example, the standardized scale uses information primarily from the calibration of θ parameters for placing IRFs on the same scale. The present results suggest that the standardized scale consistently yields larger SMSD values for the discrimination parameter when compared with the other two scales. How effective is the standardized scale, then, in practical use for determining a common scale? The answer depends on the amount of bias in $\hat{\theta}$. If a researcher is confident that this amount of bias is negligible, or that there is an efficient way of correcting bias in ability estimates, the standardized scale could still be a simple and useful tool for finding the transformation constants α and β.

If the source of calibration data is kept the same, a major change in criterion functions appears

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

to have a minor effect on the comparison of accuracy between parameter estimates. For example, both the Stocking and Lord scale and the Samejima scale use information from the calibration of discrimination and difficulty. The Stocking and Lord scale follows the least squares principle: It rescales the estimated IRFs. The Samejima scale, on the other hand, rescales both the theoretical parameters and the corresponding estimates with their first principal component.

Simulation results indicate that the Samejima and Stocking and Lord scales yield a very similar pattern of SMSD values for individual parameters, especially as the test is lengthened. A generalization of the results to other simulation situations, however, remains to be demonstrated in further research. The computation requirements for these two scales are substantial, in comparison with the standardized scale. The Gauss-Newton procedure for minimizing $F(\alpha,\beta)$ in the Stocking and Lord scale is supported on most computers. Therefore, the Stocking and Lord scale might prove to be more tractable than the Samejima scale. The SMSD equation is closer to the criterion function specified in the Stocking and Lord scale, and this fact raises some questions about the simulation results presented in this study; they might be biased in favor of the Stocking and Lord scale.

A limited simulation does not allow for a definite conclusion concerning the validity of any scale. The three scales are similar in that they all suggest that the *b* parameters are generally more accurately estimated than are the θ and the *a* parameters, and that an increase in test length will improve accuracy in calibrating the *b* and θ parameters. Because the data source used in a criterion function determines a portion of the magnitude of SMSD, knowledge concerning the standard error or bias of the data source is essential. Furthermore, the study suggests that any comparison of the accuracy between two parameter estimates should be done with reference to more than one scaling procedure.

References

Daniels, R. W. (1978). *An introduction to numerical methods and optimization techniques.* New York: Elsevier North-Holland.

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

International Mathematical & Statistical Libraries (1982). *Reference manual* (9th ed.). Houston TX: IMSL, Inc.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-form reliability. *Psychometrika, 48,* 233–245.

Samejima, F. (1984). *Comparison of the estimated item parameters of Shiba's word/phrase comprehension tests obtained by LOGIST-5 and those by the tetrachoric method* (Res. Rep. 84-2). Knoxville TN: University of Tennessee, Department of Psychology.

Samejima, F. (1986, April). *Effect of guessing parameter on the estimation of the item discrimination and difficulty parameters when three-parameter logistic model is assumed.* Paper presented at the meeting of the American Educational Research Association, San Francisco CA, U.S.A.

Scales, L. E. (1985). *Introduction to non-linear optimization.* New York: Springer-Verlag.

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

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton NJ: Educational Testing Service.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21,* 91–111.

Author's Address

Send requests for reprints or further information to Michelle Liou, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/