ACT Research Report Series 97-7

A Study of Modified-Guttman and IRT-Based Level Scoring Procedures for Work Keys Assessments

E. Matthew Schulz

Michael J. Kolen

W. Alan Nicewander

ACT

October 1997

A Study of Modified-Guttman and IRT-based Level Scoring Procedures for Work Keys Assessments

E. Matthew Schulz
Michael J. Kolen
W. Alan Nicewander

Abstract

This paper compares modified Guttman and item response theory based procedures for classifying examinees into ordered levels when each level is represented by several multiple choice test items. In the modified Guttman procedure, within-level number correct scores are mapped to binary level mastery scores. Examinees are then assigned to levels in Guttman fashion on the basis of their observed pattern of binary mastery scores over levels. In the IRT procedure, examinees are assigned to levels on the basis of their total number correct score over all items on the test. An IRT-based method is used to assess the reliability and classification consistency of both level scoring procedures. In comparison to the modified Guttman procedure, IRT-based number correct level scores are more reliable and consistent across forms. Guttman-consistent patterns of mastery over levels can also be inferred from number correct level scores.

A Study of Modified-Guttman and IRT-based Level Scoring Procedures for Work Keys Assessments

1. Introduction

The goal of this study was to evaluate, and possibly improve upon, a modified Guttman procedure for assigning examinees to levels when each level is represented by several multiple choice test items. The psychometric qualities of concern were the reliability and classification consistency of level scores, and the accuracy of inferences that examinees are masters of levels up to and including their estimated level of skill. Equivalence of level scores across forms was also investigated. In the following section (Section 2), the motivation for level based assessment and the assessment instruments and sources of data for the study are described.

Section 3 describes the modified Guttman scoring procedure applied to the data. A modified form of Guttman scaling was developed for the assessments in this study because each level was represented by more than one item. This form of Guttman scaling could conceivably be used to increase the number of observations per level and the reliability of level scores in assessments that currently rely on Guttman scaling. Traditional Guttman scales, using just one rating or binary item per level, are not uncommon in assessments based on stage theories of cognitive or physical functioning. For examples of Guttman scales see Boulton-Lewis (1987), Fox & Tipps (1995), Katz and Akpom (1976), and Lund, Foy, Sipperelle & Strachan (1984).

Section 4 presents an alternative, IRT-based procedure in which examinees are assigned to levels on the basis of their total number correct score over all items in the assessment. This procedure draws on ideas about the relationship between item response theory and a Guttman scale consisting of just one item per level (Cliff, 1983;

Andrich, 1985; Wilson, 1989a). The IRT-based procedure, also called number-correct level scoring, can conceivably be applied in other large scale assessments in which levels are defined on continuous, latent proficiency scales primarily for the purpose of describing the achievement of individuals or groups (see for example, National Assessment Governing Board, 1990; Masters, Adams, & Lokan, 1994).

Section 5 presents an IRT-based method for evaluating the reliability and classification consistency of the level scores in this study. This method is an adaptation of a general procedure outlined by Kolen, Zeng, & Hanson (1996) for estimating the conditional standard errors and reliability of any secondary scale score that is a nonlinear transformation of number correct scores. Both the modified Guttman level scores and the number correct level scores in this study are nonlinear transformations of number correct scores. The method detailed here, particularly for number correct level scores, should be of general interest. Level scores are frequently nonlinear transformations of number correct scores. These include the functional levels into which examinees are classified for course placement, licensure and certification, and personnel selection. When tests are used to classify examinees into levels or at-or-above a given level, test developers are obligated to report the reliability of the level scores, the standard error of measurement at level boundaries, and the percentage of examinees who would be classified the same way by parallel forms (*Standards*, 1985).

Results are presented in Section 6 and a discussion follows in Section 7. Some additional points and literature references are presented in the discussion because they

are better understood with reference to more specific features of the scoring procedures and to other conceptual issues presented below.

2. The Motivation and Assessment Design for Level Scoring

In the Work Keys Program at ACT, employee skills and the requirements of jobs are compared in each of several areas of skill. These include Applied Mathematics, Reading for Information, Teamwork, Applied Technology, Observation, Locating Information, Listening, and Writing. This paper is concerned only with Reading for Information and Applied Mathematics. Analyses identical to those described in this paper were performed for all skill areas represented by multiple choice test items (all but Listening and Writing).

In order to simplify the comparison of employee skills to jobs, each skill area was conceived, at the outset, as consisting of just a few discrete, ordered levels (ACT, 1997). Five levels of skill were defined for Applied Mathematics and Reading for Information. Each level was first defined by means of text and examples. The level definitions were then used in the job profiling component of the program to establish the skills and level of each skill required by a particular job. It is assumed in the job profiling that qualified employees will have mastery of all levels up to and including the level of skill identified as necessary for the job.

The assessment component of the program was therefore developed to support in some explicit fashion the notion that examinees are masters of levels up to and including their estimated level of skill. Each level of Reading for Information and Applied Mathematics was represented by a pool of multiple choice items. A panel of

experts decided that an examinee should be able to get at least 80% of the items representing a level correct in order to be classified as having mastery of the level.

Three parallel forms each of Reading for Information and Applied Mathematics were developed. These are labeled Form 1 to Form 3. (Form 1 of Reading for Information is distinct from Form 1 of Applied Mathematics.) These forms had no items in common. They contained six items per level for a total of thirty items per form and eighteen items per level across forms. Mathematics items were stand-alone. Reading for Information items were attached to passages, which were nested within level. No more than three items were attached to a single passage. A fourth form of Reading for Information (Form 4) contained the same items as Form 1, but in different format. This form is included in this study. Where only one set of statistics for a particular item was needed (specifically Equation 2 below), the item statistics from Form 4 were used.

In order to facilitate the comparability of level scores in the modified Guttman scoring procedure described in the next section, the forms were made as parallel as possible within levels. This was done on the basis of classical and IRT item statistics from pilot data. The pilot item statistics also showed that average item difficulty by level increased across levels, but that the difficulty of individual items overlapped somewhat across levels.

All test forms were administered to randomly equivalent groups of high school juniors and seniors using a spiraling process within classrooms. Raw score summary statistics are shown in Table 1. Sample sizes ranged from 1925 to 2046 per form. Mean

number correct scores ranged from 20.3 to 21.2 for Reading for Information forms and from 18.8 to 19.1 for Applied Mathematics forms. Reading for Information number

TABLE 1

Summary Statistics for Work Keys Tests

Statistic	Reading for Information				Applied Mathematics		
	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3
Raw Score Summary Statistics							
Sample Size	2032	2020	2024	1987	2022	2046	1996
Mean	20.7	21.2	20.3	20.9	18.8	19.0	19.1
S.D.	4.4	4.2	4.5	4.2	5.1	4.9	4.8
Skew	-1.12	-.96	-.83	-.91	-.26	-.38	-.53
Kurtosis	1.89	1.39	.87	1.19	-.04	-.03	.29
Raw Score Reliability							
KR_{20}	.79	.77	.80	.77	.83	.81	.80
3-PL IRT	.79	.78	.81	.78	.85	.83	.82

correct scores showed substantially more skew (approximately -1) and kurtosis (>1, except for Form 3) than those for Applied Mathematics. Reliability coefficients based on the KR_{20} formula ranged from .77 to .80 for Reading for Information, and from .82 to .85 for Applied Mathematics. Reliability estimates based on the 3-PL IRT model (Kolen, et al., 1996) were slightly higher than the KR_{20} coefficients.

3. Modified Guttman Level Scoring

Traditional Guttman scaling of binary items defines as many levels as there are items in the assessment (Guttman, 1950). A priori notions about the number of levels

that constitute a variable can be confirmed by traditional Guttman scaling only if all of the items within a level are responded to in exactly the same way--an examinee would have to get either all or none of the items within a level correct. There is no formal incorporation of measurement error in the procedures for constructing the scale or in evaluating Guttman-inconsistent patterns. Guttman scaling is therefore characterized as deterministic.

In order to preserve the a priori number of levels defined for Work Keys skills (five, for the skills in this study), as well as to accommodate in some fashion the measurement error known to exist with multiple choice test items, a binary level mastery score was defined as a function of less-than-perfect performance on the items representing a level. Guttman scoring rules were then applied to the level-mastery scores. With six items per level, within-level number correct scores ranged from 0 to 6. A binary level mastery score was defined to be 1 if the examinee got at least five of the items representing the level correct, 0 otherwise. [Five items correct (83%) was as close as one could come with six items to the 80%-correct-standard.]

To illustrate, if an examinee earned within-level number correct scores of 6, 5, 4, 4, and 3 on levels 1 through 5 respectively, the examinee received binary level mastery scores of 1, 1, 0, 0, and 0, respectively. This examinee's mastery pattern is '11000' in the notation of this study. The examinee's total number correct score, 22, was not a direct factor in the Guttman scoring, but was used in the number-correct level-scoring procedure to be described later.

Table 2 shows all of the mastery patterns that were observed for Reading for Information and Applied Mathematics. Out of thirty-two (2^5) possible patterns of

TABLE 2

Mastery Patterns and Guttman Level Scores

Mastery Pattern	Number of Errors		Guttman Level Score		Frequency[e]	
	Type A[a]	Type B[b]	NLM[c]	HCL[d]	Reading for Information	Applied Mathematics
11111	0	0	5	5	134	149
11110	0	0	4	4	1453	595
11101	2	1	4	3	91	47
11011	2	2	4	2	21	1
10111	2	3	4	1	0	1
01111	2	4	4	0	2	0
11100	0	0	3	3	2748	1808
11010	2	1	3	2	314	69
11001	2	2	3	2	18	11
10110	2	2	3	1	41	12
10101	2	3	3	1	1	1
11000	0	0	2	2	2018	1698
10100	2	1	2	1	167	154
10010	2	2	2	1	22	4
10001	2	3	2	1	3	9
01100	2	2	2	0	35	6
01010	2	3	2	0	3	0
00110	4	4	2	0	1	0
10000	0	0	1	1	600	1060
01000	2	1	1	0	70	49
00100	2	2	1	0	10	5
00010	2	3	1	0	3	2
00001	2	4	1	0	1	1
00000	0	0	0	0	307	382
					8063	6064

[a]Goodenough-Edwards errors as described in Edwards (1957).
[b]Number of score pairs with '0' on easier level, '1' on harder level (Meijer, 1994).
[c]Number of levels mastered.
[d]Highest contiguous level mastered.
[e]Frequencies are combined across test forms.

mastery scores, twenty four were observed. The last two columns show the frequency of each pattern by area of skill. Ninety percent (7260 of 8,063) of the Reading for Information patterns and 93.4% (5692 of 6064) the Applied Mathematics patterns were Guttman-consistent.

Two types of error variables are given for each mastery pattern in Table 2. Type A errors are counted according to the Goodenough-Edwards method (Edwards, 1957). The number of errors by this method equals or exceeds the number of errors by Guttman's original method of counting errors (also described in Edwards, 1957). The Goodenough-Edwards coefficients of reproducibility were .957 and .975 respectively for Reading for Information and Applied Mathematics. By these indices, which are more conservative than Guttman's coefficient, the binary Work Keys level-mastery scores appear to be highly Guttman-scalable.

The Type B error variable was created in order to evaluate IRT model fit as described at the end of Section 5. The Type B error variable is the number of Guttman errors as defined by Meijer (1994). It is the number of score pairs with a '0' on the easier level and a '1' on the more difficult level. The pattern '01111', for example, contains four such score pairs. [With five levels, there are ten possible pairs, but the number of possible Type B errors for a person who has mastered r levels is $r(5-r)$.] The percentage of patterns with one Type B error was 8% (642) for Reading for Information and 5.3% (319) for Applied Mathematics. The percentage of patterns with more than one Type B error was 2% (161) for Reading for Information and 1% (53) for Applied Mathematics.

The columns labeled "NLM" and "HCL" in Table 2 show the level scores assigned to the mastery patterns according to two scoring rules. The NLM level score is the number of levels mastered (NLM). The HCL level score is highest contiguous level mastered by the examinee. NLM and HCL-level scores are the same for Guttman-consistent patterns. For Guttman inconsistent patterns, the HCL-level score is always lower than the NLM-level score. The pattern, '11010', for example, is HCL-scored '2' and NLM-scored '3'.

HCL scoring was the modified Guttman scaling procedure used for Work Keys assessments because mastery of all levels below the highest level of mastery was deemed critical for successful job performance in the job profiling component of Work Keys. NLM scoring was included in this study because it is the traditional Guttman scoring procedure for binary data, and is therefore also of interest.

4. IRT-based Number Correct Level Scoring

IRT is regarded by some experts as a framework for Guttman scaling. Cliff (1983) remarked that a Guttman scale is one of the best examples of a good idea in all of psychometric measurement, but that is difficult to apply this idea without the concepts of a true score and an item characteristic curve. Using an example of a four-item test, Andrich (1985) formalized the relationship between a Guttman scale and the Rasch (1-parameter) model. Andrich observed that a perfect Guttman scale is unidimensional in a stronger sense than the unidimensionality of IRT models that contain just one ability parameter. He shows that in order for items to be Guttman-scalable their ICC curves on the latent proficiency scale must not cross. Number correct scores must also be a

sufficient statistic for ability. Given these requirements in an IRT framework, levels correspond to points on the latent proficiency scale, and there are exactly as many levels as there are items or binary scores in the assessment.

For levels represented by more than one item, these ideas can be modified in interesting and useful ways. Instead of a level being a point on a latent scale, a level can be a range of ability on the scale. Examinee's can be classified as masters or nonmasters of a level on the basis of their true score on the domain of items representing the level rather than on the basis of their probability correct score on an individual item. Level-mastery scores estimated on this basis will be Guttman scalable as defined by Andrich (1985), if the level-characteristic curves do not cross. Interestingly, the item characteristic curves of individual items may cross within as well as across levels, and marginal item difficulties may even overlap across levels, without causing level-characteristic curves to cross.

Levels corresponding to ranges of ability on an IRT scale have been defined in other assessments. In the National Assessment of Educational Progress, lower boundaries for Basic, Proficient, and Advanced levels were set by using a modified Angoff standard setting procedure (ACT, 1994). Masters, et al., (1994) decided on the location of level boundaries by carefully studying the spatial arrangement and content of items calibrated to the scale. Level boundaries were drawn with regard to the content of items falling into the same level. Wilson (1989a) associated the lower boundary of a level with a .8 probability of mastering the easiest item within the level, and the upper boundary with a .8 probability of mastering the hardest item in the level.

For the Work Keys assessments, level boundaries on a latent proficiency scale were defined on the basis of a percentage correct true score on level pools. Let X_i represent the random binary score on item i, where x_i=1 indicates a correct response and x_i=0 indicates an incorrect response. Unidimensional IRT models the probability of a correct response to item i in terms of a single, continuous ability parameter, θ, and one or more item parameters. Let $P_i(\theta)$ represent the conditional-on-theta probability of a correct response to item i. In the 3-item-parameter IRT model,

$$P_i(\theta) = P(X_i=1|\theta) = c_i + 1 - \frac{c_i}{1+e^{-(1.7*a_i)(\theta-b_i)}} \qquad (1)$$

where a_i, b_i, and c_i define respectively, the slope, location, and lower asymptote of the item characteristic curve, or trace line of P_i on θ. The 3-parameter model was used for Work Keys assessments in order to accommodate variation in the characteristic curves of operational items. Separate BILOG (Mislevy & Bock, 1990) analyses were performed on the data from each test form. Since forms were administered to randomly equivalent groups, all items were considered to be calibrated on a common scale.

Figures 1 and 2 show the plots of individual and mean item p-values by level. The item p-value, which is the proportion of examinees who answered the item correctly, is inversely related to item difficulty. There are eighteen individual item p-values plotted within each level. The trend lines connect the p-value means. These plots confirm the observation, based on pilot data, that item difficulties overlap across levels, but that average item difficulty increases substantially by level.

FIGURE 1. Item p-values by level of Reading for Information (18 items per level).

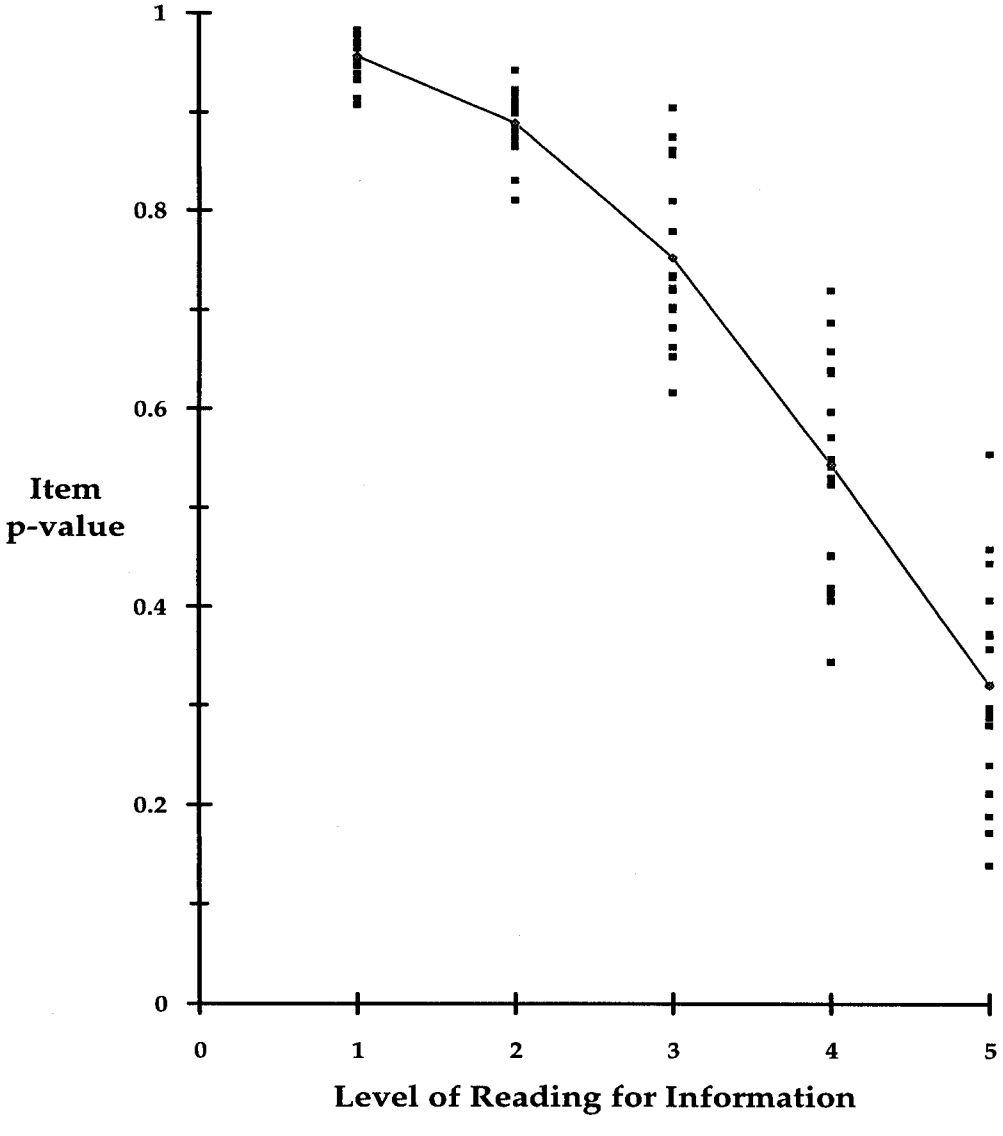
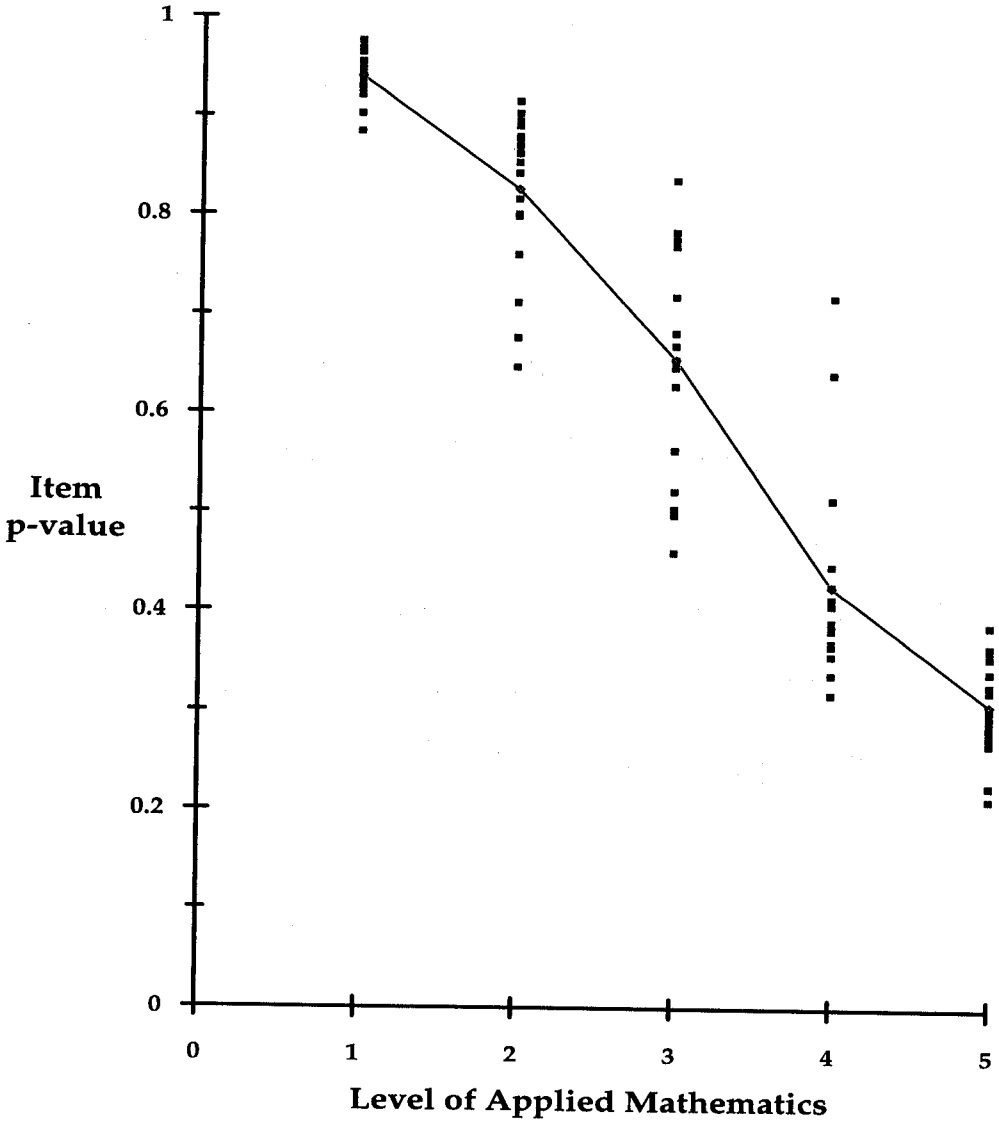


FIGURE 2. Item p-values by level of Applied Mathematics (18 items per level).



Level characteristic curves were computed as follows. Let $P_\ell(\theta)$ represent the conditional proportion correct true score on the level ℓ pool, where ℓ=1 for the easiest level and ℓ=5 for the hardest level. Let $P_{\ell i}(\theta)$ represent the probability of a correct answer to item i in the level ℓ pool, conditional on θ, and recall that there are eighteen items per level. Then,

$$P_\ell(\theta) = \frac{\sum_{i=1}^{18} P_{\ell i}(\theta)}{18}, \qquad \ell=1,\ldots,5. \tag{2}$$

Figures 3 and 4 show the level characteristic curves in terms of $P_\ell(\theta)$. The level characteristic curves are widely separated over most values of theta. Except for very low values of θ, the conditional, relative difficulty of levels is the same as the marginal relative difficulty of levels. At a .8 criterion of mastery, the mastery patterns that are 'true' among examinees are all Guttman-consistent. For any examinee, one would infer mastery of easier levels first.

FIGURE 3. Reading for Information level characteristic curves.

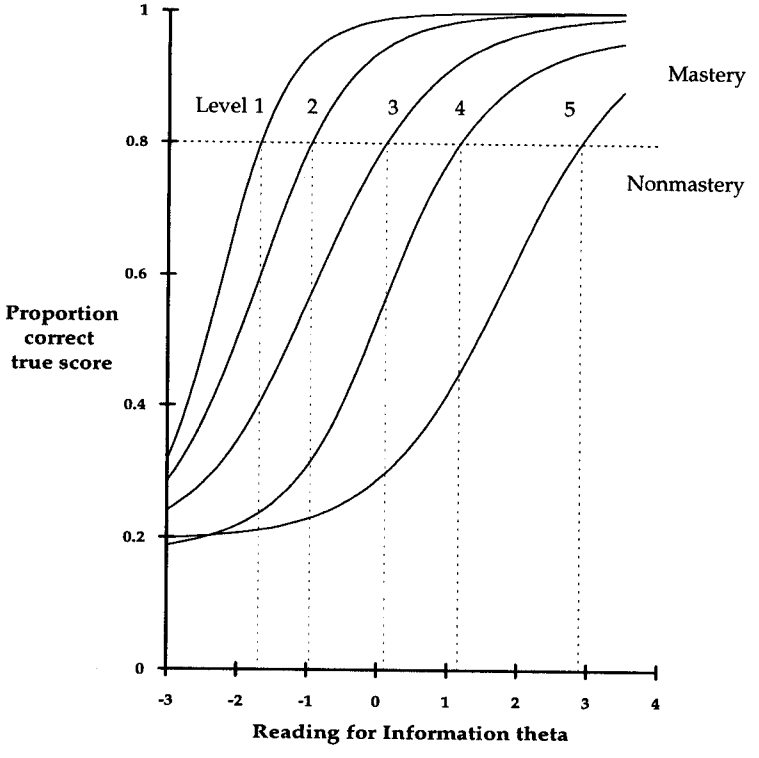
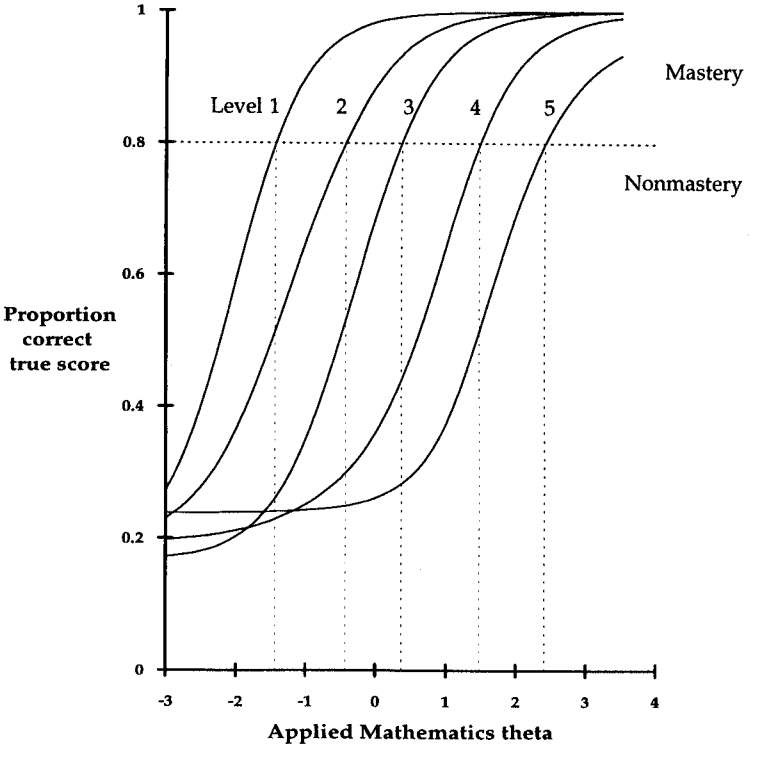


FIGURE 4. Applied Mathematics level characteristic curves.



The vertical lines in Figures 3 and 4 show the θ-coordinates that are paired with .8 proportion correct true scores on levels. These θ-coordinates are the lower boundaries for Work Keys levels on the θ-scale. Let θ_ℓ represent the lower boundary for level ℓ on the theta scale. θ_ℓ can be found by setting the left hand side of Equation 2 [$P_\ell(\theta)$] to .8 and using successive values of θ in the right hand side (by the method of half intervals) until a desired level of accuracy is reached. The columns labeled "Targeted" in Table 3 show values for θ_ℓ, ℓ=1,...,5 within each area of skill.

TABLE 3

Targeted and Operational Standards for Levels in Theta Metric

	Reading for Information					Applied Mathematics			
		Operational[b]					Operational[b]		
Level	Targeted[a]	Form 1	Form 2	Form 3	Form 4	Targeted[a]	Form 1	Form 2	Form 3
1	-1.68	-1.57	-1.72	-1.66	-1.67	-1.43	-1.43	-1.51	-1.54
2	-0.95	-1.04	-1.06	-1.06	-1.07	-0.43	-0.37	-0.47	-0.49
3	0.11	0.24	0.13	0.30	0.21	0.36	0.48	0.42	0.40
4	1.15	1.25	1.02	1.26	1.18	1.48	1.28	1.36	1.36
5	2.88	2.86	2.73	2.40	2.80	2.40	2.34	2.19	2.56

[a]Targeted thetas correspond to θ_ℓ as described in the text.
[b]Operational thetas correspond to $\theta_{f\ell}$ as described in the text.

Form-specific, number correct cutoff scores for assigning examinees to levels were then established. Let $C_{f\ell}$ represent the cutoff score for assigning examinees to level ℓ on form f, let $\theta_{f\ell}$ represent the theta upon which $C_{f\ell}$ is the form-f conditional true score, let P_{fi} represent the probability of a correct answer on i'th item within form f, and recall that there are 30 items in each form. Then

$$C_{f\ell} = \sum_{i=1}^{30} P_{fi}(\theta_{f\ell}) \ . \qquad (3)$$

Equation 3 was used iteratively to find a form f cutoff score for each level, ℓ, that made $\theta_{f\ell} \approx \theta_\ell$. Because $C_{f\ell}$ is not a continuous variable on the true score scale (it is necessarily an integer), $\theta_{f\ell}$ cannot precisely match θ_ℓ on any form and will vary across forms. The choice of cutoff scores among alternate forms for level ℓ was therefore guided by the joint goals of making $\theta_{f\ell}$ equivalent across forms as well as close to θ_ℓ.

Table 3 shows the final operational cutoff thetas by level and test form ($\theta_{f\ell}$), and Table 4 shows the number correct cutoff scores ($C_{f\ell}$) for levels by test form. The operational cutoff thetas for a given level satisfy the joint goals of being as consistent across forms as possible, and as close to the target cutoff theta as possible. Number correct cutoff scores were also quite consistent across forms within the same skill area. The number correct cutoff score for Reading for Information Level 3, for example, was twenty-two on all forms.

TABLE 4

Number-correct Cutoff Scores for Levels

	Reading for Information				Applied Mathematics		
Level	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3
1	14	14	13	14	12	12	12
2	17	17	16	17	17	17	17
3	22	22	22	22	21	21	21
4	25	25	25	25	25	25	25
5	28	28	28	28	29	28	28

Example: For Form 1 of Reading for Information, number correct scores of 0 to 13 were scored "level 0", number correct scores of 14 to 16 were scored "level 1", etc.

It should be noted that the strategy of making operational cutoff thetas as equivalent as possible across forms does not mean that the number correct cutoff scores for levels will be consistent across forms. If forms had varied much in overall difficulty, the number correct cutoff scores for a given level would also have varied. The consistency in number correct cutoff scores across forms owes to the fact that the forms had been constructed on the basis of pilot data to be similar in difficulty within levels (and therefore in average difficulty). The means of number correct scores in Table 1 show that the forms within each skill area are similar in average difficulty.

Level scores were assigned using the cutoff scores shown in Table 4. For example, examinees whose total number correct score on any form of Applied Mathematics was 0 to 11, received a level score of 0. Examinee's scoring from 12 to 16 on any form of Applied Mathematics received a level score of 1. Examinees who scored 28 on Applied Mathematics received a level score of 4 if they took Form 1, but a level score of 5 if they took Forms 2 or 3.

5. Methods of Evaluating and Comparing Level Scoring Procedures

May and Nicewander (1994) observed that Item Response Theory (IRT) contains sufficient assumptions and constructs to allow for theoretical comparisons of the psychometric properties of various scoring procedures for the same test. These authors noted that the reliability of percentile rank scores could not be predicted from the reliability of number correct scores, which can be computed by other means, such as the KR_{20}, because percentile rank is a nonlinear transform of number correct scores. From the conditional (on theta) binomial distributions of item scores under the 3-parameter

IRT model, May and Nicewander derived the conditional mean, variance, and information of number correct scores and percentile rank scores for hypothetical tests of varying degrees of difficulty and discrimination. The reliabilities of number correct and percentile scores were derived by integrating their conditional variance over the hypothetical theta distribution.

Kolen, Zeng, and Hanson (1996) outlined a general IRT-based method for estimating the psychometric properties of any secondary score scale that is a nonlinear transform of number correct scores. The basic approach, which differs from that of May and Nicewander (1994), is to estimate the compound binomial distribution of number correct scores conditional on ability (θ), via a recursion formula provided by Lord and Wingersky (1984). Kolen et al., (1996) and Kolen and Brennan (1995) describe how to implement this formula. One then applies the scale transformation to the conditional distribution of number correct scores to obtain the conditional distribution of secondary scale scores. The mean and standard deviation of this conditional secondary scale score distribution were taken as, respectively the conditional true score and standard error of the secondary scale. Conditional variance of the secondary scale score can then be integrated over the theta distribution to obtain an estimate of reliability.

This method has been successfully applied to a variety of secondary scale scores derived from various types of data, and using different IRT models. Kolen et al. (1996), used the 3-parameter model to compute the conditional standard errors and reliability of ACT scale scores. Results were comparable to those based on a strong true-score theory approach (Feldt, 1984), which was applied in the same study. Wang, Kolen &

Harris (1997) used the partial credit (Masters, 1982) and the generalized partial credit model (Muraki, 1992) model to estimate the conditional standard errors, reliability, and classification consistency of level scores derived from rating scale data. Results were satisfactory in comparison to a classical test theory-based procedure (Feldt and Qualls, 1996), which was also performed on the rating scale data.

The following procedures are stated in terms of population parameters and integrals. In implementing these procedures, estimates of item parameters and ability distributions obtained from form-specific BILOG analyses were used. The 3-parameter logistic model was specified, and Equation 1 was used to estimate the conditional item probabilities defined below. Because the posterior distribution of theta was specified in terms of fifty quadrature points, the integrals were replaced by summations. The quadrature points were equally spaced over a -4 to 4 range of θ. Computations were performed separately by form, and averaged across forms where appropriate.

The conditional probability of level mastery scores

Let $X_\ell \in (0,\ldots,6)$ represent the random number correct score on level ℓ, and let $P(X_\ell = x \mid \theta)$ represent the compound binomial distribution of X_ℓ conditional on θ. Let $U_\ell \in (0,1)$ represent the random binary level mastery score for level ℓ, and let $P(U_\ell \mid \theta)$ represent the probability $U_\ell = 1$ conditional on θ. Equation 1 was used to compute the six individual item probabilities within level ℓ, conditional on θ. These were then used in the recursive formula of Wingersky and Lord to compute $P(X_\ell = x \mid \theta)$. Then, $P(U_\ell \mid \theta) = P(X_\ell \geq 5 \mid \theta)$.

The conditional probability of NLM level scores

Let $K \in (0,1,\ldots,5)$, be a random level score, and let $P(K \mid \theta)$ represent the conditional probability distribution of level scores. In NLM scoring, K is the sum of the binary level mastery scores, so $P(K \mid \theta)$ has a compound binomial distribution which can be obtained by applying the recursive formula of Wingersky and Lord (1984) to the conditional, binomial probabilities of level mastery scores defined in the preceding paragraph.

The conditional probability distribution of HCL level scores

In HCL scoring, the level score is 0 if and only if the level mastery score for level 1 is 0. Therefore, $P(K=0 \mid \theta) = 1-P(U_1 \mid \theta)$. The level score is 1 if and only if the mastery score for level 1 is 1 <u>and</u> the level mastery score for level 2 is 0. Therefore $P(K=2 \mid \theta) = [P(U_1 \mid \theta)]*[1-P(U_2 \mid \theta)]$. And so on, as summarized in the following equation:

$$
\begin{aligned}
P(K|\theta) &= 1-P(U_1|\theta), && K=0, \\
&= \left[\prod_{\ell=1}^{K}\left[P(U_\ell|\theta)\right]\right]\left[1-P(U_{K+1}|\theta)\right], && K=1,\ldots,4 \\
&= \prod_{\ell=1}^{5} P(U_\ell|\theta), && K=5 .
\end{aligned}
\tag{4}
$$

The conditional probability distribution of number correct level scores

Let $X \in (0,1,\ldots,30)$ represent the random (total) number correct score, and let $P(X=x \mid \theta)$ represent the compound binomial distribution of X conditional on θ. Conditional item probabilities were computed using Equation 1. These were then used to compute $P(X=x \mid \theta)$ via the Wingersky and Lord algorithm. Then,

$$
\begin{aligned}
P(K|\theta) &= \sum_{x=0}^{C_{f\,1}-1} P(X=x|\theta), && K=0, \\
&= \sum_{x=C_{f\,K}}^{C_{f\,K+1}-1} P(X=x|\theta), && K=1,\ldots,4 \\
&= \sum_{x=C_{f\,5}}^{30} P(X=x|\theta), && K=5\ .
\end{aligned}
\tag{5}
$$

The conditional true score, standard error, and marginal reliability of level scores

For all scoring procedures, the conditional mean level score is:

$$
\begin{aligned}
\mu(K|\theta) &= \sum_{K=1}^{5} K*P(K|\theta)\ , \\
&=1*P(K=1|\theta) + 2*P(K=2|\theta),\ldots, + 5*P(K=5|\theta).
\end{aligned}
\tag{6}
$$

and the conditional variance of the level score is:

$$
\sigma^2(K|\theta) = \sum_{K=1}^{5} K^2P(K|\theta) - \mu^2(K|\theta)\ . \tag{7}
$$

The conditional mean is taken as the conditional true score. The square root of the conditional variance is taken as the conditional standard error. The marginal relative frequency of the level score is:

$$
P(K) = \int_{\theta} P(K|\theta)\psi(\theta)d\theta\ , \quad K=0,1,\ldots,5\ . \tag{8}
$$

where $\psi(\theta)$ is the density function of θ. The marginal mean level score is:

$$\mu(K) = \sum_{K=1}^{5} K*P(K) \ , \tag{9}$$

The marginal level score total variance is:

$$\sigma^2(K) = \sum_{K=1}^{L} K^2P(K) - \mu^2(K) \ . \tag{10}$$

The marginal level score error variance is:

$$\sigma^2(E) = \int_{\Theta} \sigma^2(K|\theta)\psi(\theta)d\theta \ . \tag{11}$$

Finally, the reliability of the level score is:

$$\rho(KK') = 1 - \frac{\sigma^2(E)}{\sigma^2(K)} \tag{12}$$

The procedures outlined above through Equation 11 were performed separately by form. The left hand side of Equations 10 and 11 were then averaged across forms. The across-form averages were used in Equation 12 to yield just one reliability coefficient per level scoring procedure within each area of skill.

Classification Consistency and Kappa Coefficients

The following procedures are modified from those of Wang, Kolen, and Lee (1993). Let $P(K'=K)$ represent the proportion of examinees who would be classified into the same skill level on two, randomly equivalent instances of measurement with a given

test form, and let $P(K' \geq \ell \cap K \geq \ell)$, $\ell=1,\ldots,5$ represent the proportion of examinees whose classifications on two random occasions of measurement with a given form would be the same with respect to being at or above a given skill level, ℓ. These probabilities may be considered conditionally on θ as well as unconditionally. Conditionally,

$$P\left[(K'=K)|\theta\right] = \sum_{K=0}^{5} P^2(K|\theta) \ , \tag{13}$$

because of local independence, and

$$P(K' \geq \ell \cap K \geq \ell)|\theta) = \left[\sum_{K=0}^{\ell-1} P(K|\theta)\right]^2 + \left[\sum_{K=\ell}^{5} P(K|\theta)\right]^2 \qquad \ell=1,\ldots,5 \ . \tag{14}$$

Unconditionally,

$$P(K'=K) = \int_{\Theta} [P(K'=K)|\theta]\psi(\theta)d\theta \ , \tag{15}$$

and,

$$P(K' \geq \ell \cap K \geq \ell) = \int_{\Theta} [P(K' \geq \ell \cap K \geq \ell)|\theta]\psi(\theta)d\theta, \quad \ell=1,\ldots,5 \ . \tag{16}$$

Indices of classification consistency may be more clearly interpreted if they are compared to the chance probabilities of correct classification using the marginal probabilities of the classification categories. Towards this end, we also computed Cohen's kappa coefficient (κ) (Cohen, 1960). The general formula for κ is:

$$\kappa = \frac{P_o - P_c}{1 - P_c} \ . \tag{17}$$

where P_o is the observed proportion of agreement and P_c is the proportion of agreement expected from the marginal probabilities under the assumption of independence. For same-level consistency, $P_o = P(K'=K)$, and

$$P_c = \sum_{K=0}^{5} P^2(K) \ . \tag{18}$$

For at-or-above level 0 consistency, $P_o = P(K' \geq \ell \cap K \geq \ell)$, and

$$P_c = \left[\sum_{K=0}^{\ell-1} P(K)\right]^2 + \left[\sum_{K=\ell}^{5} P(K)\right]^2 \quad \ell=1,\ldots,5. \tag{19}$$

Form-specific values of P_o and P_c (chance agreement) were averaged within scoring procedure and kappa (κ) was computed from the averages

Form equivalence

Form equivalence for a given level scoring procedure was assessed in part by comparing the observed frequency distributions across forms. Since forms were administered to randomly equivalent groups, and were designed to be parallel, the percentage of examinees at a given level should be relatively constant across forms within a skill area. As an additional check on form equivalence, the across-form variance of the left-hand side of Equation 6 (conditional true score) was computed and plotted as a function of θ for each procedure.

Ad hoc assessments of model fit

Number correct level scores can occasionally be inconsistent with the observed pattern of mastery. An examinee assigned to level 3 on the basis of his total score, for

example, could have a level-mastery pattern of '11000', '11010', or '11110'. All of these patterns (which include Guttman-*consistent* patterns) conflict at one or more levels with the pattern of true mastery inferred from a level score of 3. For example, the previously mentioned pattern of within-level number correct scores, '6, 5, 4, 4, 3', which is a level mastery pattern of '11000' and would be scored 2 in NLM and HCL scoring, sums to 22 and is therefore scored at level 3 in number correct scoring (for both areas of skill, see Table 4). A total number correct score of twenty-two could receive any HCL-level score from 0 to 4, and any NLM-level score from 1 to 4, depending on the examinee's pattern of within-level number correct scores across levels.

To address this issue, person fit was evaluated along the lines recommended by Mislevy and Bock (1990) for tests of fewer than ten items. Instead of computing the probability of each of 2^L person score patterns over items in an L-item test, the probability of each of the 2^5 patterns of observed level-mastery score over the five levels of each skill was computed. Let $\underline{U}$ represent a random person vector of observed level mastery scores. The binary elements of this vector are represented by U_ℓ, ℓ=1,...,5 defined previously. The marginal probability of $\underline{U}$, is:

$$P(\underline{U}) = \int_{\Theta} \prod_{\ell=1}^{L} P(U_\ell=1\,|\,\theta)^{U_\ell}(1-P(U_\ell=1\,|\,\theta))^{U_\ell}\psi(\theta)d\theta \ . \qquad \textbf{(20)}$$

Equation 20 was applied separately for each form and results were averaged across form. Observed and expected frequencies were then collapsed into three categories of departure from Guttman consistency: no departure (zero Type B errors), mild departure (one Type B error), and more serious departure (more than one Type B error). Type B

errors were described in the section on Guttman scoring. No formal statistical test was applied to the comparison of observed and expected frequencies in these categories.

6. Results

Marginal Distributions of Level Scores

The columns labeled "Obs." (for 'observed') in Table 5 and Table 6 describe the distributions of level scores (averaged across forms), that resulted from applying the level scoring procedures. Summary statistics for the level-score distributions are quite similar across scoring procedures. The mean level score in Reading for Information was 2.6 using HCL scoring, 2.7 using NLM scoring and 2.5 using number-correct scoring. The variance of Reading for Information level scores was 1.2 for both HCL and number-correct scoring, and 1.1 for the NLM procedure. The means of Applied Mathematics level scores were 2.2, 2.3, and 2.2 respectively for HCL, NLM, and number correct scoring, and the variances were 1.4, 1.3, and 1.4 respectively.

There were a few notable discrepancies in the percentages of examinees placed into a given level. Number correct scoring consistently placed more examinees into level 0 (one or two percentage points more in both areas of skill). Also with number correct scoring, the percentage of examinees at level 2 of Reading for Information was approximately ten points higher and the percentage of examinees at level 3 was approximately seven points lower in comparison to HCL and NLM scoring.

TABLE 5

Marginal Distribution and Reliability of Reading for Information Level Scores

Level Score	Level Scoring Procedure					
	Highest Contiguous Level Mastered		Number of Levels Mastered		Number-correct Level	
	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.
0	5	6	4	4	6	6
1	10	11	8	9	8	8
2	29	29	28	28	38	38
3	35	35	39	38	30	29
4	18	18	19	19	17	17
5	2	2	2	2	2	3
Mean:	2.6	2.6	2.7	2.7	2.5	2.5
Total Var:	1.2	1.3	1.1	1.2	1.2	1.2
Error Var:		.48		.39		.32
Predicted Reliability:		.62		.67		.74

Note: Numbers in upper panel of table are percentages. Obs = observed. Pred. = IRT-predicted.

TABLE 6

Marginal Distribution and Reliability of Applied Mathematics Level Scores

Level Score	Level Scoring Procedure					
	Highest Contiguous Level Mastered		Number of Levels Mastered		Number-correct Level	
	Obs.	Pred.	Obs.	Pred.	Obs.	Pred.
0	7	8	6	7	8	7
1	20	21	18	19	21	21
2	29	29	31	30	32	31
3	31	31	31	31	28	27
4	10	10	11	11	11	11
5	3	3	2	2	3	3
Mean	2.2	2.2	2.3	2.3	2.2	2.2
Total Var:	1.4	1.4	1.3	1.3	1.4	1.4
Error Var:		.38		.33		.30
Predicted Reliability		.73		.75		.78

Note: Numbers in upper panel of table are percentages. Obs = observed. Pred. = IRT-predicted.

IRT-predicted frequency distributions and summary statistics are shown in the columns labeled "Pred." in Tables 5 and 6. There are thirty-six predicted percentages of examinees by level, area of skill, and scoring procedure. Twenty two of these are in exact agreement with the observed percentages (rounded to the nearest whole number) and the remaining fourteen are within one percentage point. This level of predictive accuracy existed for each area of skill and scoring procedure.

There were six IRT-predicted means and variances of level scores (two skill areas and three scoring procedures). Rounded to the nearest tenth, the IRT-predicted mean agreed exactly with the observed mean, and the IRT-predicted variance differed by .1 from the observed variance in only two cases (HCL and NLM scoring of Reading for Information).

Marginal Error Variance and Reliability of Level Scores

The last two rows of Tables 5 and 6 show the predicted marginal error variance and reliability of level scores. Number-correct level scoring has the lowest predicted marginal error variance (.32 for Reading for Information and .30 for Applied Mathematics) and the highest predicted reliability (.74 for Reading for Information and .78 for Applied Mathematics). NLM level scores had the next lowest predicted marginal error variance (.39 and .33) and next highest predicted reliability (.67 and .75). HCL level scores had the highest predicted marginal error variance (.48 and .38) and lowest predicted reliability (.62 and .73). It should be noted that the IRT-predicted reliability of the *raw score* (.78 to .85 in Table 1) is greater than the predicted reliability of any *level score* from the same test form.

Conditional True Level Scores

Figures 5 and 6 show the conditional true level scores (averaged across forms) by level scoring procedure. Visual inspection of these plots leads to the conclusion that the conditional true level score does not vary substantially by level scoring procedure. The conditional NLM true level score slightly exceeds the conditional HCL true level score at all values of theta, as expected (see explanation of these scoring procedures above). The number correct conditional true level score has a steeper slope on θ than either of the other two procedures. At high values of theta (over 1.5) the number correct conditional true level score tends to be clearly higher than that of the Guttman procedures.

FIGURE 5. Conditional true level score in Reading for Information by level scoring method.

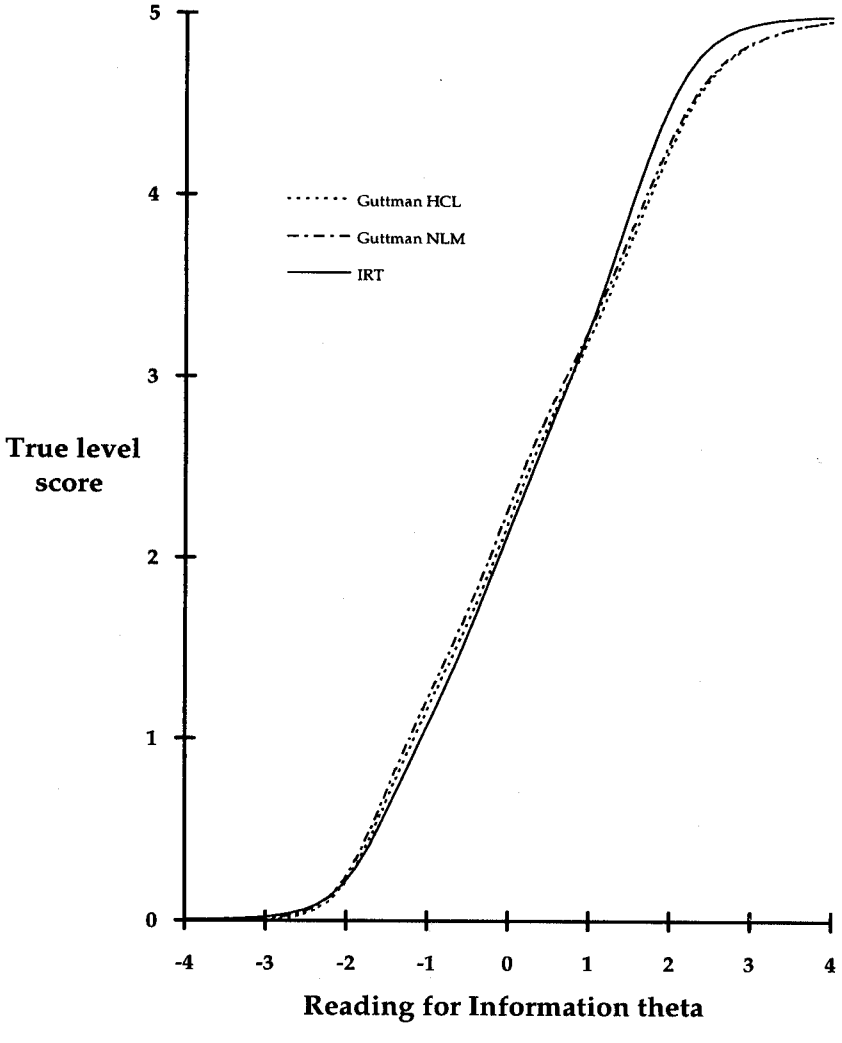
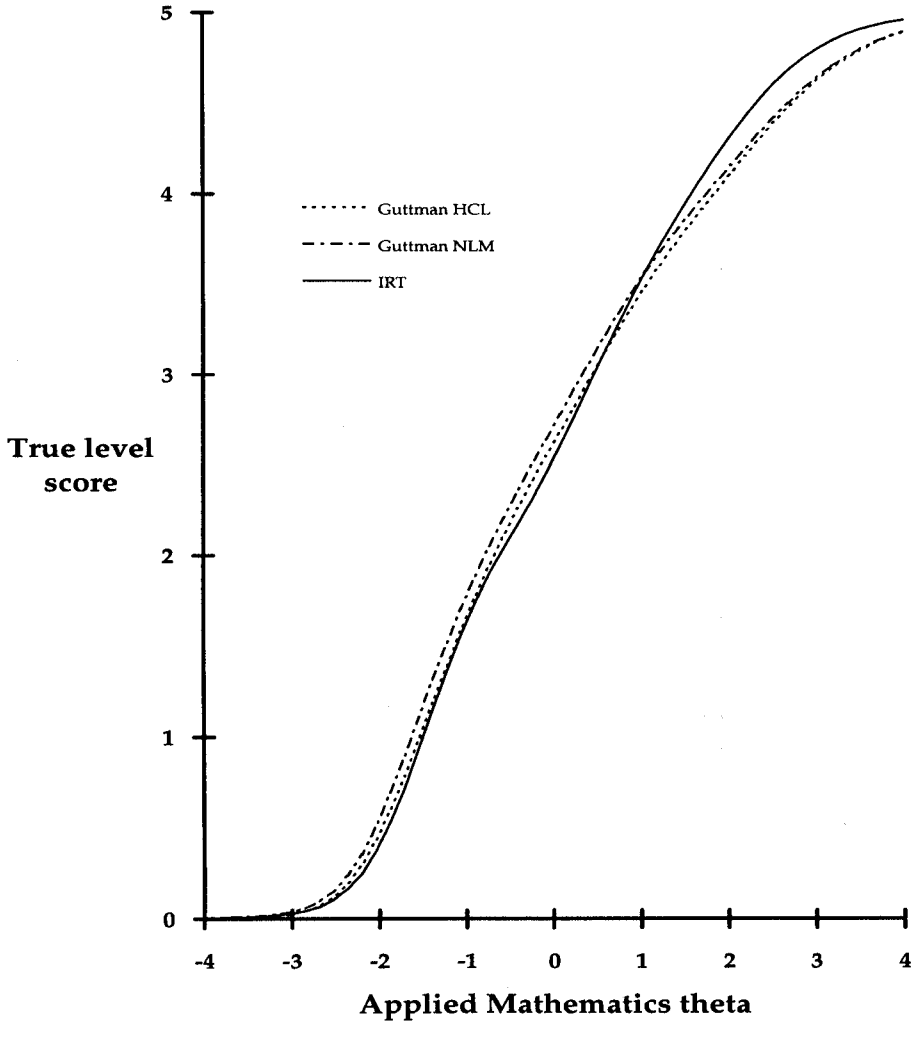


FIGURE 6. Conditional true level score in Applied Mathematics by level scoring method.



Conditional Measurement Error of Level Scores

Figures 7 and 8 show the plots of conditional measurement error by level scoring procedure. Vertical lines separate the theta scale into levels according to the targeted thetas that were established for the number correct level scoring (see Table 3). These boundaries are not directly related to NLM and HCL level scores. Nevertheless, the measurement error of all level scoring procedures tends to rise near level boundaries on the theta scale.

From visual inspection of Figures 7 and 8, it is clear that conditional error of number-correct level scores is predicted to be as low as, or lower than, that of the Guttman level scoring procedures at all values of theta. NLM level scoring is a close second, and has about the same measurement error as number correct level scoring for theta values ranging from about +.5 to 2.0. This range corresponds to all of level 3 and the lower half of level 4, in both areas of skill. HCL level scores have substantially higher conditional measurement error at all values of theta, except at the lower boundary of level 1 of Reading for Information.

FIGURE 7. Conditional standard error of level score in Reading for Information by scoring procedure.

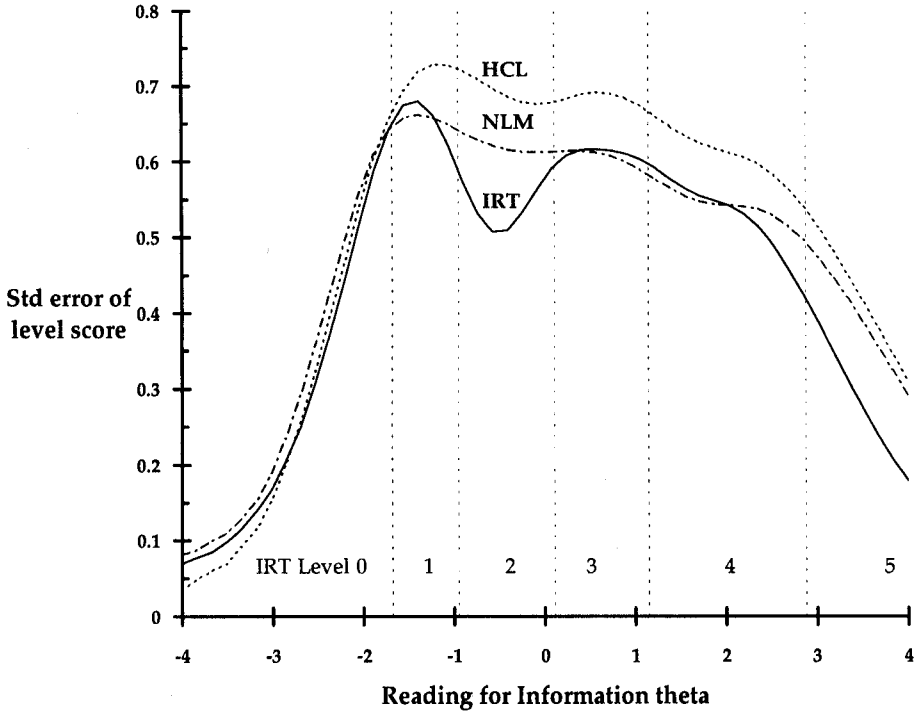
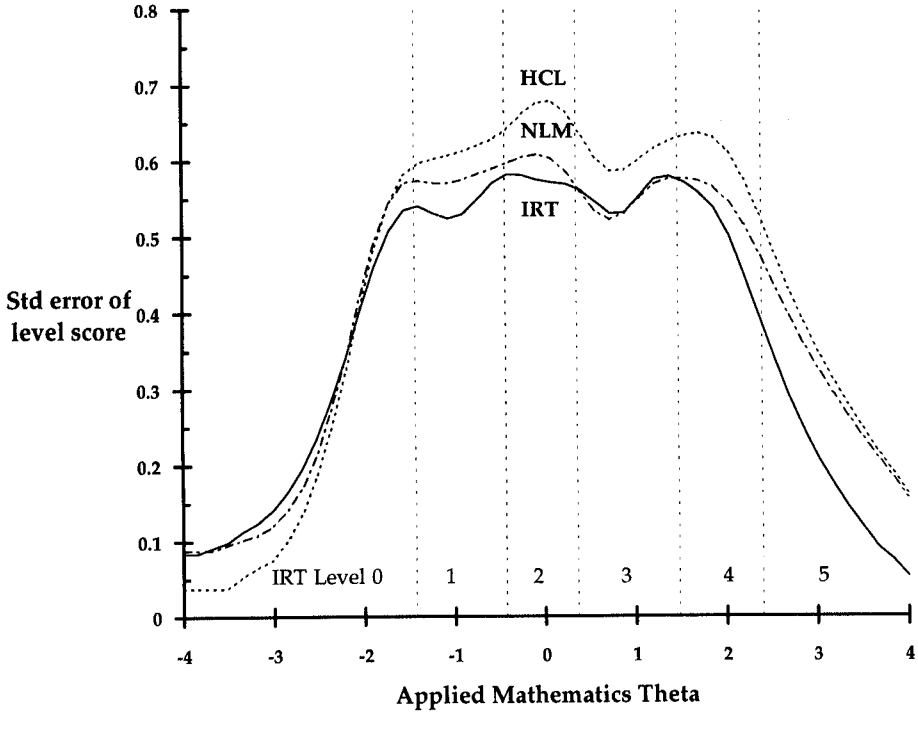


FIGURE 8. Conditional standard error of level score in Reading for Information by scoring procedure.



Classification Consistency of Level Scores

Table 7 shows the indices of predicted, "parallel form" classification consistency (P_o, P_c and κ) by skill and level scoring procedure. Compared to either of the modified Guttman procedures, number correct level scoring has generally higher indices of observed consistency (P_o) and kappa (κ). "Same Level" classifications and classifications at-or-above mid-levels (levels 2, 3, and 4) are always more consistent when number correct level scoring is used. When classifications are at-or-above extreme levels (levels 1 and 5), Guttman procedures tend to be equally or slightly more consistent than number correct level scoring in terms of P_o, but not in terms of the kappa coefficient. The kappa for number correct scoring is higher in every case except one (classifications ≥ Level 1 in Applied Mathematics), where the kappa for both NLM and number correct scoring is .56.

In terms of the 0 to 1 kappa coefficient scale, the advantage of number correct scoring over NLM scoring is as large as .09 (.56 versus .47 for ≥ Level 3 Reading for Information) and the advantage over HCL scoring is as large as .14 (.56 versus .42 for Level 3, Reading for Information).

TABLE 7

Classification Consistency of Level Scores

Type of Classification	Skill Area and Level Scoring Procedure																	
	Reading for Information									Applied Mathematics								
	Highest Contiguous Level Mastered			Number of Levels Mastered			Number-Correct Level			Highest Contiguous Level Mastered			Number of Levels Mastered			Number-Correct Level		
	P_o	P_c	κ	P_o	P_c	κ	P_o	P_c	κ	P_o	P_c	κ	P_o	P_c	κ	P_o	P_c	κ
Same Level	.43	.26	.23	.46	.27	.26	.50	.27	.32	.48	.24	.32	.50	.24	.34	.51	.23	.37
≥ Level 1	.95	.90	.52	.97	.92	.59	.96	.91	.60	.93	.86	.53	.95	.88	.56	.94	.86	.56
≥ Level 2	.86	.73	.48	.89	.77	.54	.90	.76	.58	.80	.59	.52	.83	.62	.56	.84	.59	.61
≥ Level 3	.71	.51	.42	.75	.52	.47	.78	.51	.56	.80	.51	.59	.80	.51	.60	.82	.52	.62
≥ Level 4	.81	.68	.40	.81	.67	.42	.84	.69	.48	.89	.79	.50	.89	.77	.52	.91	.77	.60
≥ Level 5	.97	.96	.20	.97	.96	.20	.96	.95	.28	.97	.95	.41	.97	.95	.41	.97	.94	.50

Note. P_o is the predicted proportion of times a student would be classified the same way on two randomly equivalent occasions of measurement with the same test form. Same level = classified into the same level both occasions. ≥ Level 1 = classified either ≥1 both occasions or <1 both occasions, etc. P_c = chance level of consistency, given marginal frequency distributions. κ = predicted kappa coefficient.

Form Equivalence of Level Scores

Tables 8 and 9 show the observed percentages of examinees at each level of skill by form and level scoring procedure. With few exceptions, the proportion at each level is reasonably consistent across forms for all procedures. The absolute difference between pairs of forms in terms of the percentage at a given level tends to be less than five percentage points regardless of scoring procedure.

Exceptions involve almost exclusively the HCL and NLM scoring of Form 3 of Reading for Information and Form 2 of Applied Mathematics. With Form 3 of Reading for Information, the percentages at levels 1 and 4 are over ten percentage points greater, and the percentage at level 3 is over ten percentage points less than with other forms. With Form 2 of Applied Mathematics, the percentage at level 2 is over five percentage points greater, and percentage at level 3 is over five percentage points less than with other forms.

TABLE 8

Percentages of Examinees by Level of Reading for Information, Form, and Level Scoring Procedure

	Highest Contiguous Level mastered				Number of Levels Mastered				Number-correct Level			
Level	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4	Form 1	Form 2	Form 3	Form 4
0	5.8	3.6	6.4	5.3	4.1	2.9	4.6	3.4	6.0	5.2	5.6	5.7
1	8.1	9.8	**14.8**	8.6	7.1	7.7	**11.2**	7.8	7.1	7.1	8.0	7.8
2	31.1	30.3	28.2	28.3	29.7	28.2	26.4	27.7	38.3	35.9	42.1	36.9
3	38.7	36.3	**25.4**	40.1	41.7	39.1	**31.9**	42.7	31.2	29.6	27.6	30.3
4	14.7	17.8	**24.0**	15.1	16.1	20.1	**24.6**	16.6	15.2	19.4	14.5	17.2
5	1.5	2.1	1.3	1.7	1.5	2.1	1.3	1.7	1.9	2.9	2.2	2.0

Note: Percentages in bold type differ substantially from those from one or more other form(s) for the same level and level scoring procedure.

TABLE 9

Percentages of Examinees by Level of Applied Mathematics, Form, and Level Scoring Procedure

Level	Highest Contiguous Level mastered			Number of Levels Mastered			Number-correct Level		
	Form 1	Form 2	Form 3	Form 1	Form 2	Form 3	Form 1	Form 2	Form 3
0	7.0	7.3	7.7	6.0	6.4	6.4	8.5	7.7	7.2
1	20.1	18.9	22.5	18.8	17.0	19.7	22.2	20.0	19.8
2	32.0	**26.2**	30.2	32.8	**27.4**	32.5	31.2	32.2	31.7
3	28.4	**35.9**	27.7	29.2	**36.7**	28.5	25.3	27.8	29.4
4	9.1	9.4	10.4	9.7	10.1	11.3	10.4	9.2	9.8
5	**3.4**	2.3	1.6	**3.4**	2.3	1.6	2.4	3.1	2.2

Note: Percentages in bold type differ substantially from those from one or more other form(s) for the same level and level scoring procedure.

Another noteworthy discrepancy is that, with Form 1 of Applied Mathematics and HCL and NLM scoring, the percentage at level 5 (3.4%) is substantially greater (proportionally) than the Form 2 (2.3%) and Form 3 (1.6%) percentages at level 5.

Number correct level scoring reduces these inconsistencies considerably. With Form 3 of Reading for Information, the percentage at level 2 (42.1) differs by only five points from the Form 2 percentage (35.9); the percentages for all other levels are within three percentage points of other forms. With Form 2 of Applied Mathematics, the percentages at all levels are within three percentage points of the other forms. With Form 1 of Applied Mathematics, the percentage at level 5 (2.4%) is close to the Form 2 (3.1%) and Form 3 (2.2%) percentages.

Figures 9 and 10 show the across-form variance of the conditional true level score. For all of the scoring procedures, this variance tends to decrease towards the middle of levels on the theta scale, and to peak at or near the level boundaries. With few

exceptions, number correct true level scores tend to be less variable across forms than the NLM and HCL true level scores. Exceptions are at the boundary between levels 2 and 3 in Reading for Information (Figure 9), and at Applied Mathematics theta values corresponding to level 1 and lower (Figure 10).

FIGURE 9. Across-form variance of conditional true level score in Reading for Information by level scoring procedure.

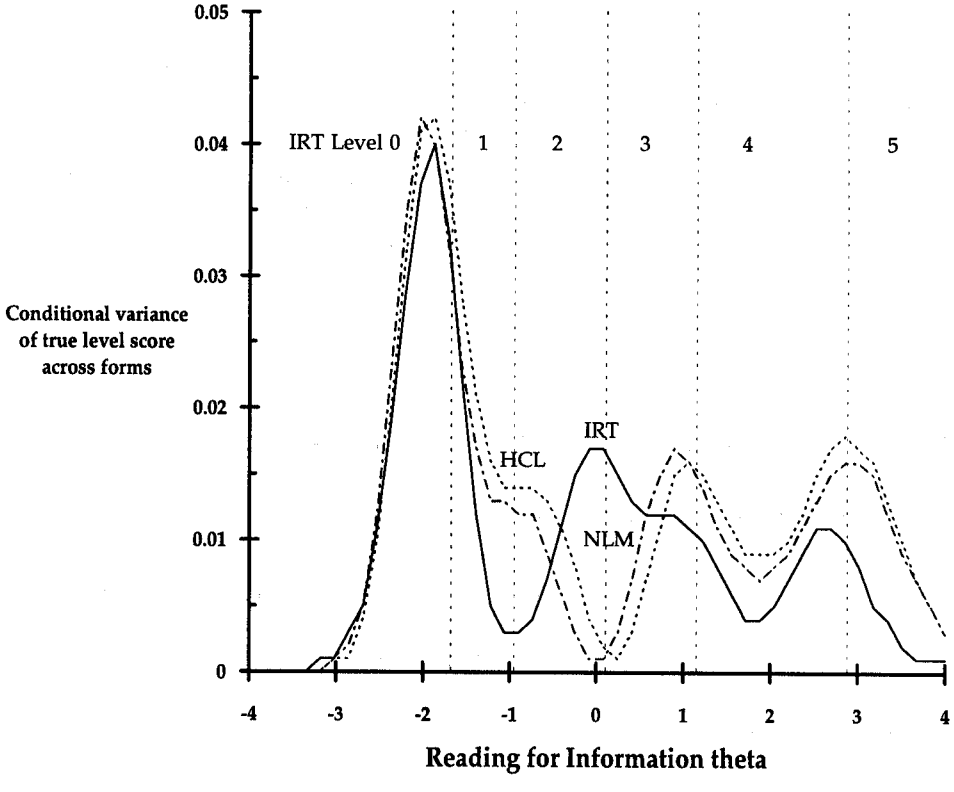
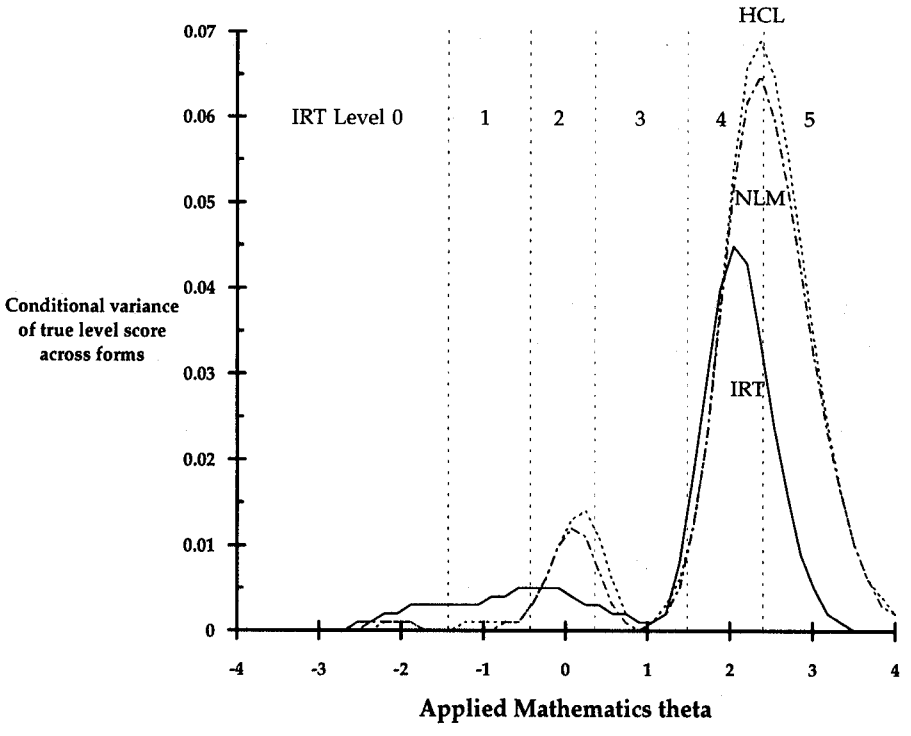


FIGURE 10. Across-form variance of conditional true level score in Reading for Information by level scoring procedure.



IRT Model Fit

The predicted percentages of mastery patterns in Applied Mathematics, by category of Guttman-consistency were: 93.4% Guttman consistent, 5% one Type B error, and 1% more than one Type B error. These predictions were within one percentage point of the observed percentages: 93.7% Guttman consistent, 5.3% one Type B error, and 1% more than one Type B error.

Mastery patterns in Reading for Information were somewhat less Guttman consistent than predicted: Percentages for Guttman consistency were 92% predicted versus 90% observed; one Type B error was 7% predicted versus 8% observed; more than one type B error was 1% predicted versus 2% observed.

7. Discussion

Number correct level scoring is predicted to be the most reliable and consistent level scoring procedure used in this study. The measurement error of number correct level scores was approximately equal to or lower than that of modified Guttman level scores at all points on the latent ability scale. In Applied Mathematics, number correct level scores were more reliable than NLM level scores by .03 and more reliable than HCL level scores by .05. In Reading for Information, number correct level scores were more reliable than NLM level scores by .07 and more reliable than HCL level scores by .12. Kappa coefficients for same-level and at-or-above level classifications by number-correct level scoring exceeded those for NLM level scores by as much as .09, and exceeded those for HCL level scores by as much as .14.

Form equivalence was also more satisfactory with number correct level scoring than with the Guttman style procedures. This was evident in the comparison of observed frequency distributions across forms and in the variance of form-specific, conditional true scores.

We have some confidence in these predictions because the 3-PL model on which the predictions are based adequately represents related, key aspects of these data. For all level scoring procedures, there was a good match between IRT-predicted and observed marginal means, variances, and frequency distributions of level scores (Table 5 and Table 6). The 3-PL model also predicted the percentages of observed mastery patterns shown in Table 2 reasonably well. In Applied Mathematics, Guttman-inconsistent patterns occurred no more frequently than predicted by the model. In

Reading for Information, there were only slightly more Guttman-inconsistent patterns than predicted.

This evidence of model fit is consistent with the notion that for these assessments, levels do not represent distinct factors or dimensions of achievement, at least for the practical task of predicting mastery of levels. In comparison to the within-level number correct score, the total number correct score contains more information about the examinee on the common dimension, and should therefore predict performance on any subset of items more accurately. Discrepancies between a level mastery score and the mastery inferred from the number-correct score can therefore be attributed to error in the observed mastery score.

Reading for Information passage effects, which are nested within level, probably inflated the rate of Guttman-inconsistent patterns from that predicted by a unidimensional IRT model. But passage effects, being nested within forms also, are not properly modeled either for inferring mastery of across-form level pools from form-specific level mastery scores. It is therefore unclear which scoring procedure is less adversely affected by passage effects. Since only 1% of the observed mastery patterns in Reading for Information were both 'seriously' Guttman-inconsistent (more than one Type B error) and unaccounted for by the IRT model, number correct level scores were judged to be a reasonable basis for inferring mastery of levels in this area of skill.

Assuming the data fit a unidimensional model, the reliability results can be explained informally by considering that reliability increases with the amount of information about differences between examinees. Given two, congeneric measures of

a trait, the more reliable measure is the one that contains more information about differences between examinees. Number correct levels, which range from 0 to 5, clearly represent the same trait as the number correct score, which ranges from 0 to 30, but clearly contain less information about differences between examinees. The reliability of number correct levels (Tables 5 and 6) is therefore less than the reliability of number correct scores (Table 1).

The reliability of level scoring procedures likewise depends upon the amount of information (about examinee differences) represented in the scores upon which the level score is based. A number correct level depends upon the total number correct score, which is the sum of within-level number correct scores. NLM level scores depend upon the total number of levels mastered, which is the sum of level-mastery scores. A level mastery score contains less information about an examinee's performance than the within-level number correct score. Therefore NLM level scores are based upon less information about examinee differences, and are therefore less reliable, than number correct level scores. HCL level scores are least reliable because they represent the sum of information (as represented by binary level-mastery scores) only up to the highest level of nonmastery. An HCL score of 0, for example, contains information about an examinee's performance on only the first (easiest) level.

The greater form-equivalence of number correct level scores stems from the fact that the total number correct score scale is more finely graded than the within-level number correct score scale. The finer grading allows one to more closely approximate with any given test form, a fixed standard such as the targeted thetas in Table 3, and

therefore to make alternate forms more equivalent with respect to a fixed standard. A more finely graded scale also permits finer adjustments, if any are required, for differences among forms in the difficulty of items representing levels.

A review of indices for evaluating Guttman scales (Cliff, 1983) includes several for estimating parallel forms reliability. Cliff described the most successful of these as modifications of the KR_{18} formula, and recommends that they receive wider consideration. Like the coefficient of reproducability, however, these indices could have been applied only to the binary level-mastery scores in this study. They could not have been used to assess the reliability of the number correct level scores. The Kolen et al. (1996) method was uniquely applicable to all of the level scoring procedures used in this study. This procedure, in combination with the use of several items per level, provided a unique perspective on Guttman scales.

Guttman scaling is not uncommon in assessments where development is perceived to occur in qualitative jumps along a cumulative scale of discrete levels. It is unclear, however, whether the Guttman scalability of data in these areas is due to the discrete nature of levels or to features that can be accounted for within a stochastic, unidimensional framework. If levels are widely separated on a unidimensional scale, for example, or if items are highly discriminating, one can achieve a good Guttman scale, as shown by the binary level scores in this study. The coefficient of reproducability for the binary level mastery scores in this study exceeded .96. But the estimated parallel forms reliability of a Guttman (NLM) scale constructed from these level-mastery scores is only .75 for Applied Mathematics and .67 for Reading for Information--not as high as

the predicted reliability of number-correct level scores (.78 and .74 respectively). These results depend on various conditions, including the number of items per level and the separation of levels on the latent ability scale.

Other results in this study point to the important conclusion that number correct level scores have the same meaning level scores were intended to have under Guttman scoring procedures. Figures 3 and 4 show that a Guttman pattern of mastery, based on eighty-percent correct criterion true scores, can be inferred from number correct level scores. Figures 5 and 6 show that the conditional, true level score was not substantially changed by number correct level scoring. The discrepancies that exist between modified Guttman and number correct true levels at high values of theta in Figures 3 and 4 can be explained by the lower measurement error of number correct level scores as follows.

The level score is a bounded variable, and therefore its conditional (error) distribution will tend to show progressively greater skew as the mean (true score) approaches its upper (negative skew) and lower (positive skew) limits. Distributions with relatively less (error) variance will tend to show less skew, and the mean (true score) will therefore be relatively closer to the limit. The number correct true level score is, compared to HCL and NLM scores, relatively lower at low values of theta and higher at high values of theta.

There were no surprising or unacceptable discrepancies between level scoring procedures in terms of the conditional and marginal distributions of level scores. Compared to HCL scoring, Number correct level scoring did not change the marginal mean and variance of level scores to any appreciable extent (Tables 5 and 6). This was

important because HCL-level scoring had been used in the operational scoring of Work Keys, and some users had established longitudinal trends of summary statistics based on HCL scoring.

Number-correct level scoring tended to place more examinees at level 0, but this could be a closer reflection of the true score distribution, given the lower conditional measurement error of number correct levels. As shown in Figures 7 and 8, number correct level scores have lower conditional measurement error near the lower boundary of level 1. It is also possible that number correct level scoring penalizes examinees who do not reach all of the items on the test. Post hoc analyses of omit rates, however, indicated that the tests were not speeded.

It is more difficult to explain why number correct level scoring, in comparison to HCL and NLM scoring, places quite different percentages of examinees into levels 2 and 3 of Reading for Information. From Table 3, one can see that the operational width of level 2 on the Reading for Information theta scale tends to be wider, and the operational width of level 3 narrower on most forms, than the targeted width of these levels. The targeted level 2 width, for example, is 1.06 units on the theta scale (.11 - -.95), but the Form 1 operational width is 1.28 units (.24 - -1.04). These discrepancies could explain to some degree why level 2 tends to contain more examinees and level 3 fewer than HCL and NLM levels. Also, from visual inspection of Figures 5 and 6, the trace line of the conditional, number correct true level score crosses the HCL and NLM trace lines over the level 3 range on the theta scale. This would contribute to minor discrepancies

in the marginal distributions of level scores, such as the tendency of number correct level scoring to place relatively more examinees into level 2 than into level 3.

Action and Recommendations

Based upon these results and the results of similar analyses performed on data from Work Keys assessments in other areas of skill, number correct level scoring was adopted for the Work Keys multiple choice assessments. Other areas of skill that were assessed by multiple choice items were Applied Technology, Locating Information, Observation, and Teamwork. The number of levels per skill area ranges from four to five. The number of items per level on a test form currently ranges from six to nine. The total number of items per form ranges from thirty to thirty six. Alternate forms have been developed within each area of skill. New forms are currently being developed using a variety of equating designs and techniques.

Particular attention is being given to IRT equating techniques (e.g., Stocking and Lord, 1983). In the present study, levels are defined in terms of the theta scale and the targeted thetas in Table 3. Cutoff scores were established with two objectives in mind: 1) representing the targeted standard (targeted theta) and 2) making operational standards as equivalent across forms as possible. It is conceivable that the first objective is best achieved by calibrating the new test form onto the theta scale defined in this study (or transforming the targeted theta onto the scale of the new form) and selecting cutoff scores on the new form by balancing the two objectives listed above, as done in this study. Alternative methods in which new forms are equated to a base form would perpetrate the operational standards of the base form, not the targeted standards.

Certain conditions that did not apply to the Work Keys data could favor the modified Guttman scoring procedure. If it had been practicable to represent levels with a larger number of items, modified Guttman scoring would have been more reliable and number correct level scoring might not have yielded a significant improvement in reliability. Modified Guttman scoring might have been preferred in this case because it supports inferred, Guttman-consistent patterns of mastery more clearly and directly than the abstract concepts of true scores and level characteristic curves associated with number correct level scores. Work Keys assessments were originally conceived as consisting of fifteen items per level, but this number was reduced due to operational constraints. A modified form of Guttman scoring might also be preferred to number correct level scoring if examinees at low levels of skill could not reach or be meaningfully engaged by items representing higher levels of skill. IRT solutions such as targeted testing, or post hoc fit analyses would also be appropriate for this problem, however.

Assessments involving truly discrete levels or discontinuous growth may also be best conducted along the lines of Guttman scaling or with psychometric models for discontinuity. IRT models (Wilson, 1989b) and other stochastic models (Falmagne, 1989) for discontinuity in cognitive growth have been developed. Developmental models based on Guttman scaling (Collins, Cliff, & Dent, 1988) are also available. The procedure for Guttman scaling in this study would also be appropriate when examinees are to be classified into fundamentally discrete levels. The unique characteristic of this procedure is to define level-mastery as a function of performance on more than one item per level.

The procedure for defining number correct levels in this study may also be more generally useful. Policies governing the definition of achievement levels in NAEP are consistent with the a priori assignment of items to levels (National Assessment Governing Board, 1990), as was done for Work Keys skills. Items may also be assigned to levels in the process of setting standards for achievement levels (Stone, 1995, 1996). In these cases, item difficulties and characteristic curves are likely to overlap across levels, as indicated by Figure 1 and Figure 2 in this paper. Guttman-consistent patterns of mastery can still be inferred over these levels, however, if levels are defined in terms of a common percentage correct true score on level pools. Guttman patterns of mastery, along with descriptions of levels accompanied by exemplar items, could prove highly effective in communicating test results to users. For a contrasting perspective on this form of test score interpretation, see Forsyth (1991).

To the degree assessment data do fit a model of continuous, unidimensional ability, however, level scores may be suboptimal for purposes other than describing test results to layusers. None of the level scores in this study were as reliable as the number correct scores, whose IRT-predicted reliability ranged from .82 to .85. Number correct scores (or secondary scale scores having an equal or greater number of possible values) are therefore likely to have higher correlations with external measures (i.e. validity coefficients) and to be more sensitive to change over time. Fortunately, in the Work Keys assessments, it is possible to supply more finely graded scale scores for these purposes.

As a general summary for assessments other than Work Keys, assessments that currently rely on Guttman scaling may be improved by departing from the limitations of one item per level in traditional Guttman scaling. If levels are truly discrete, then the modified Guttman scoring procedures described in this study should be considered. If data fit a model of continuous unidimensional ability, one should consider assigning examinees to levels on the basis of their total number correct score over all items in the assessment. In either case, levels can be defined in terms of a less-than-perfect standard of performance on level-specific pools of items, and Guttman-consistent patterns of mastery may be inferred from level scores. The number correct scoring procedure developed for this particular application may be useful in other assessments where levels correspond to ranges of achievement on a continuous scale. The method for estimating the psychometric characteristics of level scores in this study can be applied whenever level scores are a function (or functions) of number correct scores over all items or subsets of items and data show reasonable fit to an IRT model.

References

ACT Inc. (1994). *Setting Achievement Levels on the 1994 National Assessment of Educational Progress in Geography and U.S. History and the 1996 National Assessment of Educational Progress in Science.* Iowa City, IA: Author.

ACT Inc. (1997). *Work Keys Preliminary Technical Handbook.* Iowa City, IA: Author.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Tuma (Ed.), *Sociological Methodology* (pp 33-80). Jossey-Bass.

Boulton-Lewis, G. M. (1987). Recent cognitive theories applied to sequential length measuring knowledge in young children. *British Journal of Educational Psychology, 57,* 330-342.

Cliff, N. (1983). Evaluating Guttman Scales: Some old and new thoughts. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederick M. Lord.* Hillside, NJ.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(3), 37-46.

Collins, L. M., Cliff N. & Dent, C. W. (1988). The longitudinal Guttman simplex: A new methodology for measurement of dynamic constructs in longitudinal panel studies. *Applied Psychological Measurement, 12*(3), 217-230.

Edwards, A. L. (1957). *Techniques of Attitude Scale Construction.* Englewood Cliffs, NJ. Printice-Hall.

Falmagne, J. (1989). A latent trait theory via a stochastic learning theory for a knowledge space. *Psychometrika, 54*(2), 283-303.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44,* 883-891.

Feldt, L. S. & Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33,* 141-156.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational and Psychological Measurement: Issues and Practice, 10*(3), 3-9.

Fox, J. E. & Tipps, R. S. (1995). Young children's development of swinging behaviors. *Early Childhood Research Quarterly, 10,* 491-504.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. A. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and Prediction* (pp 60-90). Princeton: Princeton University Press.

Katz, S., & Akpom, C. A. (1976). A measure of primary sociobiological functions. *International Journal of Health Services, 6*(3), 493-507.

Kolen, M. J. & Brennan, R. L. (1995). *Test Equating: Methods and practices.* New York: Springer-Verlag.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129-140.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453-461.

Lund, M., Foy, D., Sipprelle, C. & Strachan, A. (1984). The combat exposure scale: A systematic assessment of trauma in the Vietnam War. *Journal of Clinical Psychology, 40*(6), 1323-1328.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Masters, G. N., Adams, R. & Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research, 21,* 595-609.

May, K. & Nicewander, W. A. (1994). Reliability and information functions for percentile ranks. *Journal of Educational Measurement, 31*(4), 313-325.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4), 311-314.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3. Item analysis and test scoring with binary logistic models* (2nd ed.). Mooresville, IN: Scientific Software.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

National Assessment Governing Board (1990). *Setting appropriate achievement levels for the National Assessment of Educational Progress: policy framework and technical procedures.* Washington DC: Author.

Standards for Educational and Psychological Testing. (1985). Washington, DC. American Psychological Association.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Stone, G. E. (April, 1995). *Objective Standard Setting*. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA.

Stone, G. E. (April, 1996). *The Construction of Meaning: Replicating Objectively Derived Criterion-Referenced Standards*. Paper presented at the annual meeting of the American Educational Research Association. New York, NY.

Wang, T., Kolen, M. J., & Lee, S. (1993). *Assessing inter-form consistency and equivalency based on IRT parameters*. Unpublished manuscript. American College Testing, Iowa City, Iowa, 52243.

Wang, T., Kolen, M. J & Harris, D. J. (1997). Conditional standard errors, reliability, and decision consistency performance levels using polytomous IRT. In D. J. Harris (Ed.) *Reliability Issues with Performance Assessments: A Collection of Papers*. ACT Research Report 97-3. ACT Inc. Iowa City, IA.

Wilson, M. (1989a). A comparison of deterministic and probabilistic approaches to measuring learning structures. *Australian Journal of Education, 33*(2), 127-140.

Wilson, M. (1989b). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.