Psychological Assessment
2008, Vol. 20, No. 1, 55–62

1040-3590/08/$12.00 DOI: 10.1037/1040-3590.20.1.55

Effects of Estimation Methods on Making Trait-Level Inferences From Ordered Categorical Items for Assessing Psychopathology

Levent Dumenci
Virginia Commonwealth University

Thomas M. Achenbach
University of Vermont

In assessments of attitudes, personality, and psychopathology, unidimensional scale scores are commonly obtained from Likert scale items to make inferences about individuals' trait levels. This study approached the issue of how best to combine Likert scale items to estimate test scores from the practitioner's perspective: Does it really matter which method is used to estimate a trait? Analyses of 3 data sets indicated that commonly used methods could be classified into 2 groups: methods that explicitly take account of the ordered categorical item distributions (i.e., partial credit and graded response models of item response theory, factor analysis using an asymptotically distribution-free estimator) and methods that do not distinguish Likert-type items from continuously distributed items (i.e., total score, principal component analysis, maximum-likelihood factor analysis). Differences in trait estimates were found to be trivial within each group. Yet the results suggested that inferences about individuals' trait levels differ considerably between the 2 groups. One should therefore choose a method that explicitly takes account of item distributions in estimating unidimensional traits from ordered categorical response formats. Consequences of violating distributional assumptions were discussed.

Keywords: measurement models, trait score estimation, Likert scales, psychopathology

Test scores play a crucial role in psychological assessment. This is particularly true for assessing emotional and behavioral problems in clinical and research settings. Test scores are often obtained from standardized instruments to help make decisions regarding diagnoses, treatment planning, interventions, treatment effectiveness, and outcome. Several methods are available for estimating test scores. For over a century, psychometric research has provided insights into how best to estimate test scores. However, clashes between different theoretical orientations and the highly technical literature make it hard for applied researchers and clinicians to select appropriate methods for estimating levels of psychopathology. This study aimed to compare and contrast the effects of different estimation methods on trait-level inferences for Likert scale items that assess psychopathology.

Originally proposed by Likert (1932), items with ordinal responses (e.g., strongly disagree to strongly agree) are among the most widely used response formats in attitude, personality, and psychopathology assessment in which there are no "correct" or "incorrect" responses (Spector, 1992). This contrasts with items commonly used in educational measurement. Factor analysis (FA) and related methods (principal component analysis, or PCA), item response theory (IRT), and classical test theory (CTT) are by far the most widely used frameworks for making trait-level inferences from Likert scale items that are intended to measure a unidimensional construct. This study aims to provide an answer to a practical question: Does it really matter whether we use IRT, FA, PCA, or CTT to estimate the level of an attribute (e.g., depression) from Likert scale items presumably measuring a unidimensional construct? If the answer is no, differences between IRT, CTT, FA, and PCA may have little relevance to researchers' and clinicians' trait-level inferences about individuals. If the answer is yes, however, then we need to know which method performs best and what happens if we use a less-than-optimal method for estimating test scores.

Models Originating From Item Response Theory

For estimating a latent trait from ordinal items, the most commonly used IRT models include the partial credit model (IRT-pcm; Masters, 1982) and graded response model (IRT-grm; Samejima, 1969). The IRT-pcm may be best described as an adjacent-category item response model, whereas the IRT-grm is a cumulative item response model. The IRT-pcm is also an ordinal categorical extension of the Rasch model for binary items. In terms of item parameters, the IRT-pcm estimates difficulty (or threshold) parameters for ordinal items. The model assumes that items do not differentially relate to the latent trait. Items are allowed to relate to a latent trait differentially in the IRT-grm. As a result, the IRT-grm estimates item difficulty and item discrimination parameters.

The IRT-pcm is closely related to Rasch models. In fact, the IRT-pcm "has all standard Rasch model features" (Embretson & Reise, 2000; p. 105). Wright (1999) summarized these model features as "the only laws of quantification that define objective measurement, determine what is measurable, decide which data

Levent Dumenci, Department of Social and Behavioral Health, Virginia Commonwealth University; Thomas M. Achenbach, Department of Psychology and Department of Psychiatry, University of Vermont.

Thomas M. Achenbach is president of and is remunerated by the nonprofit Research Center for Children, Youth, and Families, which publishes the Child Behavior Checklist, the Adult Self-Report, and the Older Adult Behavior Checklist and which supported this research.

Correspondence concerning this article should be addressed to Levent Dumenci, Department of Social and Behavioral Health, Virginia Commonwealth University, PO Box 980149, Richmond, VA 23298. E-mail: ldumenci@vcu.edu

are useful, and expose which data are not" (p. 70). It is important to remember that these measurement properties do not necessarily apply to any other IRT model, including the IRT-grm, where discrimination parameters are allowed to vary across items.

Items with ordered categorical response formats commonly used in assessments of attitudes, personality, and psychopathology are problematic for all IRT models because they often fail to fit the model (i.e., item responses are inconsistent with the model). Even if we are willing to sacrifice the unique measurement features of one-parameter IRT models by adding more parameters (e.g., three-parameter logistic IRT model), responses to binary and ordered categorical psychopathology items often remain inconsistent with IRT models (Reise, 1999; Reise & Henson, 2003; Reise & Waller, 2003).

Principal Components Analysis and Factor Analysis

Normal theory maximum-likelihood factor analysis (FA-ml; Joreskog, 1969) and principal components analysis (PCA; Hotelling, 1933) are often used to estimate trait levels of individuals from Likert scale items. The FA-ml assumes that item distributions are multivariate normal. However, the normality assumption is inapplicable to binary and ordered categorical variables (i.e., items), by definition. Factor analysis per se does not assume normality. Rather, it is the assumption of a particular maximum-likelihood method used to estimate the parameters of a factor model. There are other estimation methods in factor analysis that do not make the normality assumption. These methods are collectively known as asymptotically distribution-free estimators (ADF; e.g., diagonally weighted least squares with standard errors and mean- and variance-adjusted chi-square test statistic, or FA-wlsmv; Muthén & Muthén, 2004). That is, we can estimate individuals' trait levels from factor analysis using an ADF estimation method without assuming normality of binary and ordered categorical item distributions.

Derivations of principal components do not require multivariate normality assumptions. Yet, PCA does not distinguish ordered categorical items from continuously distributed items. Principal components are expressed as linear combinations of observed variables. Trait estimates from PCA are therefore differentially weighted observed variables. In contrast, traits are latent variables in factor analysis. In personality research, Goldberg and Digman (1994) have argued that FA and PCA would yield trait estimates so similar that any difference between the two would be trivial. To counter such arguments, Widaman (2004) has called for eliminating PCA from the field of psychometrics, citing its statistical inferiority to FA. Two decades ago, Cliff (1987) gave the following description of a war between FA and PCA proponents that continues today: "Some authorities insist that component analysis is the only suitable approach, and that the common-factors methods just superimpose a lot of extraneous mumbo jumbo, dealing with fundamentally unmeasurable things . . . . Militant common-factorists insist that components analysis is at best a common factor analysis with some error added and at worst an unrecognizable hodgepodge of things from which nothing can be determined" (p. 349).

Classical Test Theory

Summative scale scores are the primary way to estimate trait levels in classical test theory (CTT-sum; Lord & Novick, 1968). The CTT-sum (or a unit-weighted sum of Likert scale items) is perhaps the most common method for estimating trait levels in assessment of psychopathology (Spector, 1992). Similar to PCA, the CTT-sum does not distinguish ordered categorical items from continuously distributed items.

IRT tradition portrays CTT and its simple-sum method in particular as the major obstacle to solving measurement issues in applied psychology. For example, Embretson (1996) portrayed IRT as representing "the new rules of measurement," arguing that the CTT-sum represents the old rules of measurement. In applied psychology, it is often customary to justify the choice of an IRT model by citing limitations of CTT-sum. On the other hand, consistency between item responses and a one-factor model (or the first principal component) is often used to justify using CTT-sum based on the argument that the difference between the unit-weighted and differentially weighted item scores is trivial. Under most conditions, a linear model comprising equally weighted predictors is just as good at predicting real-world criteria in a cross-validation sample as a model comprising predictors whose weights are more precisely estimated from the original sample (Dawes, 1979). In fact, equal weights often yield a higher correlation with a criterion in a cross-validation sample than do weights estimated from the original sample (Wainer, 1976).

Purpose of the Present Study

As the end product of a measurement model, test scores indicate the extent to which individuals have certain attributes. The primary purpose of this study was to test similarities and differences between test scores obtained from different measurement models. We used three real data sets for this purpose. Measurement models included IRT (i.e., IRT-pcm and IRT-grm), CTT (i.e., CTT-sum), FA (FA-ml and FA-wlsmv), and PCA. Theoretical similarities and differences, parametric associations, and performance under ideal (e.g., large samples) and less-than-optimal (e.g., slightly misspecified models) conditions are beyond the scope of this study, as each of these topics would require several volumes.

Method

Samples

Three nationally representative samples were used in this study: (a) 2,029 6- to 18-year-olds (child sample), (b) 2,020 18- to 59-year-olds (adult sample), and (c) 489 59- to 96-year-olds (older adult sample). The child, adult, and older adult samples were originally used to construct scales for scoring the Child Behavior Checklist (CBCL; Achenbach & Rescorla, 2001), Adult Self-Report (ASR; Achenbach & Rescorla, 2003), and Older Adult Behavior Checklist (OABCL; Achenbach, Newhouse, & Rescorla, 2004), respectively. The child sample was 47.1% girls and had a mean age of 12.0 years (*SD* = 3.5). The CBCL was filled out by mothers (72.2%), fathers (21.9%), or others who knew the child well (e.g., grandparents, 5.9%). The adult sample was 58.9% women and had a mean age of 39.1 years (*SD* = 12.0). The older adult sample was 47.0% women and had a mean age of 72.3 years (*SD* = 8.2). The OABCL was filled out by a spouse or partner (45.4%), grown child (20.4%), or others who knew the older adult well (e.g., friends, 34.2%). Manuals for the instruments provide detailed descriptions of the participants, as well as reliability and validity data (Achenbach & Rescorla, 2001, for the CBCL; Achenbach & Rescorla, 2003, for the ASR; and Achenbach et al., 2004, for the OABCL).

Instruments and Procedure

The CBCL, ASR, and OABCL are standardized instruments for obtaining reports of emotional and behavioral problems, as well as adaptive functioning. For this study, we used the Aggressive Behavior scale of the CBCL (18 items; e.g., *Gets in fights, Destroys own things*, and *Threatens others*), the Anxious/Depressed scale of the ASR (18 items; e.g., *Lonely, Worries*, and *Self-conscious*), and the Functional Impairment scale of the OABCL (11 items; e.g., *Soiling, Lacks energy*, and *Clumsy*). On all three forms, each item is rated on the following 3-point scale: 0 = *not true (as far as you know);* 1 = *somewhat or sometimes true;* and 2 = *very true or often true*. Frequency distributions of all items appear in Table 1.

Data Analysis

To assess the unidimensionality of the scales, we applied the following fit indices to a one-factor model using the WLSMV method: chi-square statistic with Type I error rate set to .01, the root-mean-square error of approximation (RMSEA; Browne & Cudeck, 1993), and the Tucker-Lewis index (TLI; Tucker & Lewis, 1973). A nonsignificant chi-square test indicates a perfect fit between the model and data, which means that the inconsistency between the model and data is due only to sampling variability. However, the chi-square test is often found to be significant for large samples even if the items measure a unidimensional trait in the population. RMSEA values of less than .05 (Yu & Muthén, 2002), less than .06 (Hu & Bentler, 1999), and less than .08 (Browne & Cudeck 1993; Marsh, Hau, & Wen, 2004) have been proposed as criteria for good model fit. TLI values of greater than .90 (Browne & Cudeck 1993; Marsh, Hau, & Wen, 2004) and greater than .95 (Hu & Bentler, 1999; Yu & Muthén, 2002) have also been proposed as criteria for good model fit.

IRT-pcm, IRT-grm, FA-wlsmv, FA-ml, PCA, and CTT-sum were used to estimate the level of psychopathology for each of the three scales. The PARSCALE computer program (v. 4.1; Muraki & Bock, 2003) was used to estimate trait scores for the two IRT models (i.e., IRT-pcm and IRT-grm). We used the Expected A Posteriori (EAP) scoring algorithm (i.e., the Bayes method) for both IRT models (Bock & Aitkin, 1981). The Mplus computer program (v. 4.1; Muthén & Muthén, 2004) was used to estimate trait scores for the two factor analytic models (i.e., FA-wlsmv and FA-ml). Mplus also uses the Bayes method to estimate factor scores (Muthén & Muthén, 2004, pp. 47–48). SPSS (v. 14.0; 2006) was used to estimate principal components with the regression scoring method. SPSS sum function was also used to calculate the unit-weighted total score (i.e., CTT-sum).

We used scatter plots to visually inspect the form of relations (e.g., linear, *S*-shaped) between pairs of trait estimates from different methods within each sample. Polynomial curve fitting procedures (i.e., linear, quadratic, cubic) were used to test associations between scores estimated from different methods.[1] The changes in R^2 values from curve fitting procedures were used to test nonlinearity between scores obtained from different procedures.

Table 1
Item Frequency Distributions

Form/scale and item	0	1	2
CBCL/AGG			
3. Argues	34.7	47.0	18.3
16. Mean	87.0	12.2	0.8
19. Demands attention	61.0	29.2	10.0
20. Destroys own things	89.1	9.2	1.7
21. Destroys others' things	90.4	8.4	1.2
22. Disobedient at home	59.5	36.7	3.8
23. Disobedient at school	78.1	19.4	2.5
37. Fights	91.1	7.4	1.5
57. Attacks people	94.4	4.9	0.7
68. Screams	83.1	14.3	2.6
86. Stubborn	49.4	42.5	8.1
87. Mood changes	68.0	27.9	4.1
88. Sulks	79.8	17.4	2.8
89. Suspicious	90.5	8.4	1.1
94. Teases	68.9	25.5	5.6
95. Temper	68.6	24.7	6.7
97. Threatens	95.0	4.2	0.8
104. Loud	78.5	17.9	3.6
ASR/AXD			
12. Lonely	59.7	33.8	6.5
13. Confused	79.0	18.5	2.5
14. Cries	84.2	13.0	2.8
22. Worry future	28.8	47.7	26.5
31. Fears doing bad	83.8	14.3	1.9
33. Unloved	89.2	9.6	1.2
34. Out to get	88.4	10.2	1.4
35. Worthless	83.7	14.7	1.6
45. Nervous	52.4	40.0	7.6
47. Lacks self-confidence	62.3	32.4	5.3
50. Fearful	73.7	22.3	4.0
52. Too guilty	82.1	15.4	2.5
71. Self-conscious	52.8	38.5	8.8
91. Thinks suicide	95.9	3.5	0.6
103. Sad	72.6	24.1	3.3
107. Can't succeed	73.2	20.8	5.9
112. Worries	44.5	40.0	15.5
113. Worries opposite sex	73.3	20.9	6.1
OABCL/FI			
3. Things done	60.1	28.0	11.9
10. Dependent	80.4	15.3	4.3
16. Sits around	73.4	17.0	9.6
29. Trouble meals	81.8	13.3	4.9
54. Poor performance	87.1	9.6	3.3
55. Clumsy	85.4	11.9	2.7
68. Sleeps more	78.0	15.5	6.5
92. Lacks energy	57.0	34.4	8.6
104. Can't dress	94.1	3.9	2.0
106. Can't bathe	92.6	4.9	2.5
111. Soiling	93.0	5.5	1.4

Note. Table entries are percentages. Items are abbreviated from the versions that appear on the CBCL, ASR, and OABCL. CBCL = Child Behavior Checklist; AGG = Aggressive Problems; ASR = Adult Self-Report; AXD = Anxious/Depressed; OABCL = Older Adult Behavior Checklist; FI = Functional Impairment.

[1] As noted by a reviewer, mathematical functions such as logarithmic and exponential functions may characterize the associations between scoring methods better than polynomials do.

Table 2
Fit Indices for One-Factor Model

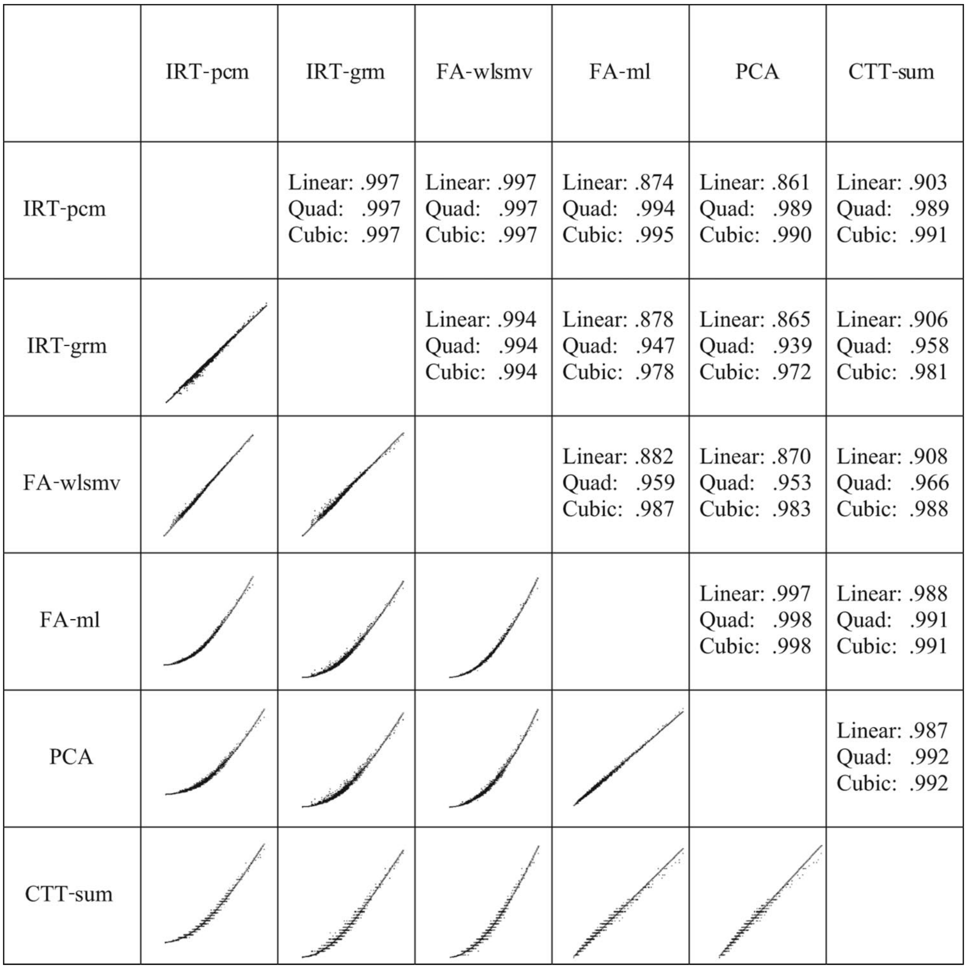
Form	Scale	Age range in years	N	Number of items	χ^2 (df)	p	RMSEA	TLI
CBCL	AGG	6-18	2,029	18	920 (86)	< .01	.069	.964
ASR	AXD	18-59	2,020	18	633 (96)	< .01	.053	.979
OABCL	FI	60-96	489	11	41 (25)	*n.s.*	.036	.993

Note. RMSEA = root-mean-square error of approximation; TLI = Tucker–Lewis index; CBCL = Child Behavior Checklist; AGG = Aggressive Behavior; ASR = Adult Self-Report; AXD = Anxious/Depressed; OABCL = Older Adult Behavior Checklist; FI = Functional Impairment.

Results

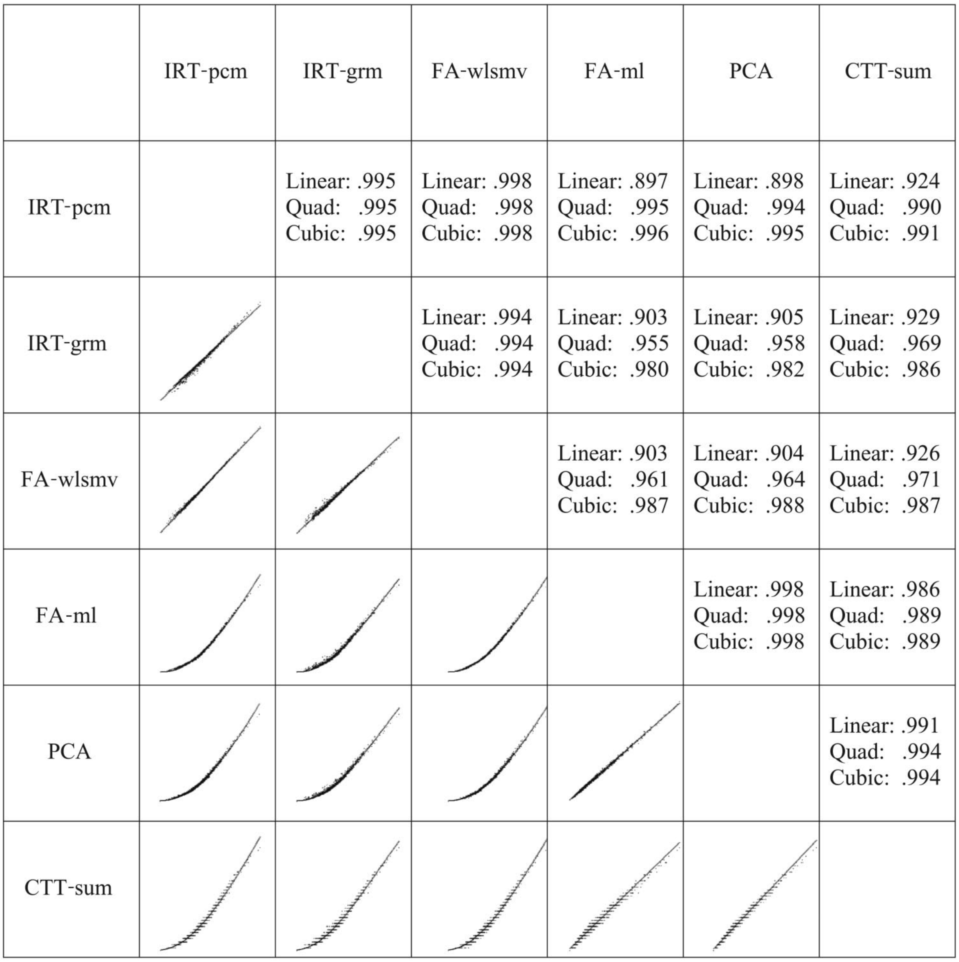
Model fit indices from FA-wlsmv appear in Table 2. Overall, results showed no major departure from unidimensionality in any of the three scales. Fit indices fell within the acceptable range. Despite the similarities in sample size and the number of items, the ASR Anxious/Depressed scale had somewhat better fit than did the CBCL Aggressive Behavior scale. The OABCL Functional Impairment scale obtained the best fit.

Scatter plots of trait estimates from the six different methods appear in the lower diagonals of Figures 1, 2, and 3 for the child, adult, and older adult samples, respectively. In terms of similarities and differences between trait estimates, visual inspection of the



	IRT-pcm	IRT-grm	FA-wlsmv	FA-ml	PCA	CTT-sum
IRT-pcm		Linear: .997 Quad: .997 Cubic: .997	Linear: .997 Quad: .997 Cubic: .997	Linear: .874 Quad: .994 Cubic: .995	Linear: .861 Quad: .989 Cubic: .990	Linear: .903 Quad: .989 Cubic: .991
IRT-grm			Linear: .994 Quad: .994 Cubic: .994	Linear: .878 Quad: .947 Cubic: .978	Linear: .865 Quad: .939 Cubic: .972	Linear: .906 Quad: .958 Cubic: .981
FA-wlsmv				Linear: .882 Quad: .959 Cubic: .987	Linear: .870 Quad: .953 Cubic: .983	Linear: .908 Quad: .966 Cubic: .988
FA-ml					Linear: .997 Quad: .998 Cubic: .998	Linear: .988 Quad: .991 Cubic: .991
PCA						Linear: .987 Quad: .992 Cubic: .992
CTT-sum						

Figure 1. Comparison of scoring methods: Aggressive Behavior scores of the Child Behavior Checklist. IRT-pcm = partial credit model from the item response theory; IRT-grm = graded response model from item response theory; FA-wlsmv = factor analysis using diagonally weighted least squares with standard errors and mean- and variance-adjusted chi-square test statistic; FA-ml = factor analysis using maximum likelihood; PCA = principal component analysis; CTT-sum = summative scale scores in classical test theory.



	IRT-pcm	IRT-grm	FA-wlsmv	FA-ml	PCA	CTT-sum
IRT-pcm		Linear: .995 Quad: .995 Cubic: .995	Linear: .998 Quad: .998 Cubic: .998	Linear: .897 Quad: .995 Cubic: .996	Linear: .898 Quad: .994 Cubic: .995	Linear: .924 Quad: .990 Cubic: .991
IRT-grm			Linear: .994 Quad: .994 Cubic: .994	Linear: .903 Quad: .955 Cubic: .980	Linear: .905 Quad: .958 Cubic: .982	Linear: .929 Quad: .969 Cubic: .986
FA-wlsmv				Linear: .903 Quad: .961 Cubic: .987	Linear: .904 Quad: .964 Cubic: .988	Linear: .926 Quad: .971 Cubic: .987
FA-ml					Linear: .998 Quad: .998 Cubic: .998	Linear: .986 Quad: .989 Cubic: .989
PCA						Linear: .991 Quad: .994 Cubic: .994
CTT-sum						

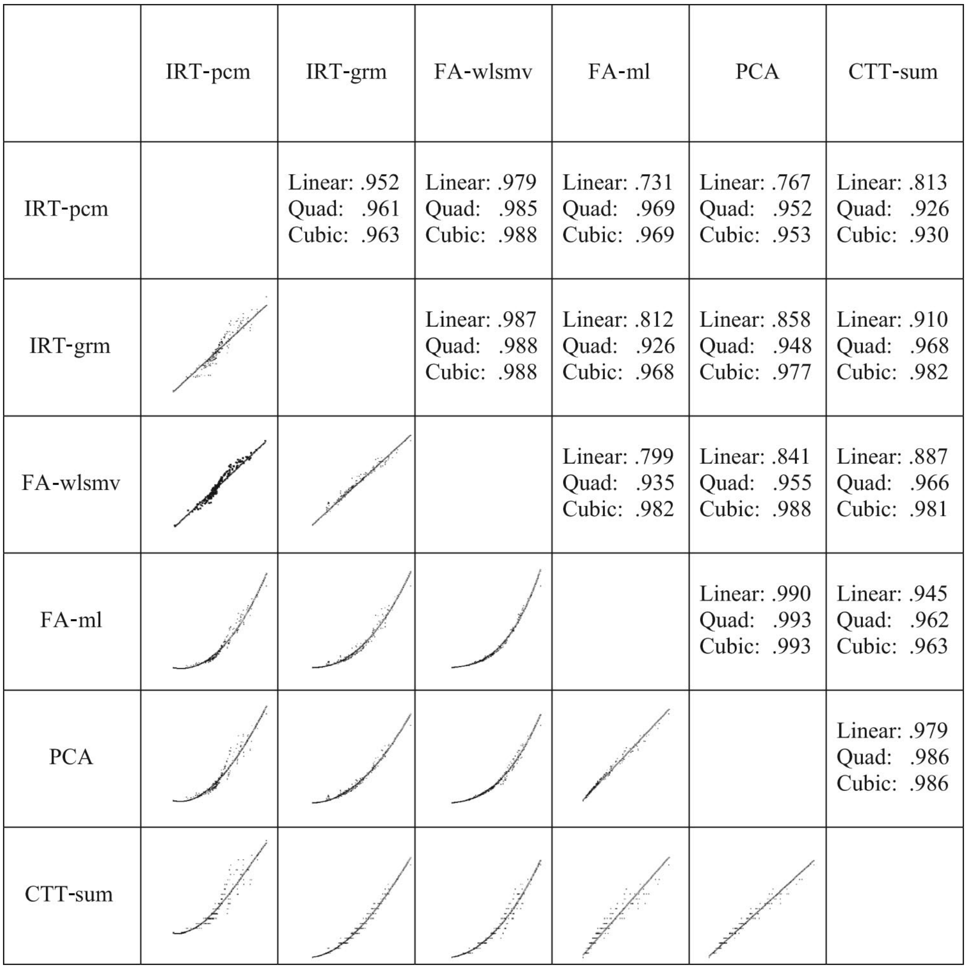
Figure 2. Comparison of scoring methods: Anxious/Depressed scores of the Adult Self-Report. IRT-pcm = partial credit model from the item response theory; IRT-grm = graded response model from the item response theory; FA-wlsmv = factor analysis using diagonally weighted least squares with standard errors and mean- and variance-adjusted chi-square test statistic; FA-ml = factor analysis using maximum likelihood; PCA = principal component analysis; CTT-sum = summative scale scores in classical test theory.

scatter plots revealed two types of curves between scores estimated from different methods: linear and curvilinear (i.e., quadratic and cubic). Two groups of methods appeared to be linearly related in all three samples. The first group consisted of IRT-pcm, IRT-grm, and FA-wlsmv, whereas the second group consisted of FA-ml, PCA, and CTT-sum. Bivariate scatter plots showed that scores fell along a straight line within each group, indicating trivial differences between scores estimated from different procedures within each group of procedures. However, nonlinear relations were found between scores obtained from the first and second groups of methods. These results were consistent in all three samples.

To quantify the findings in the scatter plot analyses, we used curve fitting procedures to test the forms of relations between different scoring methods. For the child and adult samples, scores fell on a straight line with mean R^2 value of .996 (range: .994–.998) in the group comprising IRT-pcm, IRT-grm, and FA-wlsmv. For the older adult sample, the difference in R^2 values for the quadratic curve was .001 between the IRT-grm and FA-wlsmv methods. The nonlinear relation between IRT-pcm and IRT-grm was similar to the nonlinear relation between IRT-pcm and FA-wlsmv, with the difference in R^2 value being less than 1%.

The second group of methods consisted of FA-ml, PCA, and CTT-sum. The average R^2 value for the straight line was .985 (range: .945–.998) among the three methods in the three samples. Fitting a quadratic curve increased the mean R^2 value to .989, a gain of less than 0.5%. Within the second group of methods, the strongest linear association was between FA-ml and PCA, with an average R^2 value of .995 (range: .990–.998).

Bivariate comparisons between each method from the first group and each from the second group revealed that linear associations explained a mean of 87.2% of the variance (range: .731–.929) in test scores across the three samples. The quadratic fit increased the mean R^2 to .963 (range: .926–.995), a sizable increase of 9.9%. Cubic fit improved the mean R^2 to .981, an increase of 1.8% from the quadratic fit.



	IRT-pcm	IRT-grm	FA-wlsmv	FA-ml	PCA	CTT-sum
IRT-pcm		Linear: .952 Quad: .961 Cubic: .963	Linear: .979 Quad: .985 Cubic: .988	Linear: .731 Quad: .969 Cubic: .969	Linear: .767 Quad: .952 Cubic: .953	Linear: .813 Quad: .926 Cubic: .930
IRT-grm			Linear: .987 Quad: .988 Cubic: .988	Linear: .812 Quad: .926 Cubic: .968	Linear: .858 Quad: .948 Cubic: .977	Linear: .910 Quad: .968 Cubic: .982
FA-wlsmv				Linear: .799 Quad: .935 Cubic: .982	Linear: .841 Quad: .955 Cubic: .988	Linear: .887 Quad: .966 Cubic: .981
FA-ml					Linear: .990 Quad: .993 Cubic: .993	Linear: .945 Quad: .962 Cubic: .963
PCA						Linear: .979 Quad: .986 Cubic: .986
CTT-sum						

Figure 3. Comparison of scoring methods: Functional Impairment scores of the Older Adult Behavior Checklist. IRT-pcm = partial credit model from the item response theory; IRT-grm = graded response model from item response theory; FA-wlsmv = factor analysis using diagonally weighted least squares with standard errors and mean- and variance-adjusted chi-square test statistic; FA-ml = factor analysis using maximum likelihood; PCA = principal component analysis; CTT-sum = summative scale scores in classical test theory.

Discussion

In estimating an individual's location on a continuum, does it matter if one uses scores estimated from an IRT model, factor model, principal components, or simply a sum of ordinal raw scores? To answer this question, we used six different methods to estimate scores from three scales for scoring psychopathology. Two groups of methods emerged from the analyses. The first group consisted of methods that explicitly take account of ordered categorical item distributions: IRT-pcm, IRT-grm, and FA-wlsmv. Scores estimated from these methods are strongly and linearly related to each other such that any difference among the three methods would be considered trivial.

The second group consisted of methods that do not take account of ordered categorical item distributions (FA-ml, PCA, and CTT-sum). Like the first group of methods, the methods in the second group are strongly and linearly related to each other. That is, scores estimated from these three methods are interchangeable.

This study addressed the relations between two groups of methods, an issue that has frequently been contested in the literature. Results showed that the associations between scores estimated from methods that explicitly take account of ordinal item distributions (e.g., IRT-grm) versus methods that either violate (FA-ml) or do not take account of item distributions (CTT-sum, PCA) are strongly nonlinear. The strong associations among the scoring methods indicate that the rank orders of individuals are similar across the six methods. For example, the scatter plots of scores estimated from IRT-grm and CTT-sum fall onto a straight line, when scores are transformed into rank orders. Consequently, one plausible conclusion is that choices among these methods are inconsequential because the rank ordering of individuals remains about the same across methods. However, the nonlinear associations imply that score estimation methods are not robust to violations of assumptions about underlying item distributions. Thus, this study lends support to the assertion that the CTT-sum is biased at the ends of score distribution (Wright, 1999; p. 70). Given the strong linear associations between test scores estimated from the CTT-sum, FA-ml, and PCA, this bias is present not only in raw scores (i.e., CTT-sum), but also in scores obtained from methods that ignore the ordered categorical item distributions (i.e., FA-ml and PCA).

The nonlinearity implies that the distortion of test scores is more pronounced toward the ends of trait distributions. This has clinical consequences, especially when important decisions are based on extremely low or high scores (e.g., referrals to mental health services, diagnostic decisions). For Likert-type items, mental health professionals are well advised to estimate test scores using IRT-pcm, IRT-grm, or FA-wlsmv methods. The use of these procedures will likely reduce true-positive and false-negative rates in diagnostic decisions based on estimated test scores because cut scores are often located toward the ends of trait distributions. A particular choice among these three methods appears to be inconsequential in terms of clinical utility.

There are at least three instances where ignorance of the nonlinearity has been shown to be detrimental to research. First, scores estimated from methods that do not explicitly take account of ordinal item distributions lead to inaccurate inferences from longitudinal measurement models (Fraley, Waller, & Brennan, 2000). Second, violations of distributional assumptions lead to premature rejection of measurement structures in confirmatory factor analytic studies. For example, Hartman et al. (1999) rejected a measurement structure based in part on methods that violated nonnormal item distributions (i.e., FA-ml), whereas subsequent studies supported the same structure using methods that explicitly took account of nonnormal item distributions (i.e., FA-wlsmv; Dumenci, Erol, Achenbach, & Simsek, 2004; Ivanova, Achenbach, Dumenci, Rescorla, Almqvist, Weintraub, et al., 2007; Ivanova, Achenbach, Rescorla, Dumenci, Almqvist, Bathiche, et al., 2007; Ivanova, Achenbach, Rescorla, Dumenci, Almqvist, Bilenberg, et al., 2007). Third, Embretson (1996) convincingly showed that results based on methods that violated item distributions result in spurious interaction effects in general linear models such as ANOVA.

Whereas IRT trait estimates are widely used in practical applications, trait estimates from factor analyses are often avoided due to the factor score indeterminacy issue. Given the strong linear associations between trait scores estimated from IRT-grm, IRT-pcm, and FA-wlsmv methods reported in this study, concerns over the factor score indeterminacy issue may not be justified in estimating unidimensional trait scores when ordered categorical item distributions are taken into account during the estimation process. However, simulation studies are needed to compare the true trait levels with the trait scores estimated from different measurement models to assess the detrimental effects of factor score indeterminacy. By analyzing real data, we conclude that the factor score indeterminacy issue does not differentially affect trait estimates from the IRT-grm, IRT-pcm, and FA-wlsmv methods, regardless of their true magnitude.

Do IRT and FA yield the same results? The short answer is, “it depends.” We showed that test scores estimated from IRT and FA are essentially the same when the factor analytic estimation method takes account of the correct item distributions. Earlier studies were supportive of this conclusion (McDonald, 1999; Muraki & Engelhard, 1985; Takane & de Leeuw, 1987; Wilson, Wood, & Gibbons, 1991). Muthén and Muthén (2004) took a further step by labeling FA-wlsmv for ordered categorical items as a “two-parameter logistic item response theory (IRT) model” (i.e., IRT-grm; p. 45). When assumptions about underlying distributions do not hold for ordered categorical items (e.g., multivariate normality assumption in FA-ml), however, factor analytic trait estimates are biased. The bias gets larger toward both ends of score distributions.

Do FA and PCA yield the same results? Again, the short answer is, “it depends.” Factor analytic estimates are as biased as estimates from PCA when FA does not take account of ordered categorical item distributions. However, this particular source of bias is avoidable by choosing an estimation method that correctly accounts for nonnormal item distributions.

What if a one-parameter IRT model (e.g., IRT-pcm, Rasch models) does not fit? Does it mean that the construct is “unmeasurable”? In a series of papers, Reise and colleagues convincingly showed that one-parameter IRT models typically do not fit data from ordered categorical attitude, personality, and psychopathology items (Reise, 1999; Reise & Henson, 2003; Reise & Waller, 2003). One plausible solution is to do a better job of selecting items to remove inconsistencies between item responses and the one-parameter IRT models. However, this resolution may be impractical because the possible items for measuring psychopathology constructs (e.g., anxiety) are more limited than those for measuring constructs such as mathematical ability. Another alternative is to declare that attitude, personality, and psychopathology constructs are unmeasurable. However, psychologists may not be willing to view these constructs as unmeasurable. Resorting to summing items (i.e., CTT-sum) may seem like a simple solution, but it invites measurement inaccuracies, especially in both tails of the distributions. A compromise solution could be to use an IRT or FA model that does take account of ordered categorical item distributions (e.g., IRT-grm or FA-wlsmv). It is important to emphasize that this study simply touched upon consequences of selecting various methods. It is therefore up to the users to decide what price they are willing to pay for selecting a particular method.

Generalizability of the results would be greatly enhanced with a simulation study, whereby a completely crossed design is used to provide random samples from particular populations and then scores are estimated from different methods under different conditions (e.g., sample size, strength of item–trait relations). Our results were based on only three scales, each of which came from a different sample. Yet both the strength and form of relations found between scores estimated from different methods were consistent with the measurement literature. It is therefore reasonable to expect that well-designed simulation studies would replicate our findings.

References

Achenbach, T. M., Newhouse, P. A., & Rescorla, L. A. (2004). *Manual for the ASEBA older adult forms & profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Achenbach, T. M., & Rescorla, L. A. (2003). *Manual for the ASEBA adult forms & profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of EM algorithm. *Psychometrika, 46,* 443–459.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cliff, N. (1987). *Analyzing multivariate data.* San Diego: Harcourt Brace Jovanovich.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Dumenci, L., Erol, N., Achenbach, T. M., & Simsek, Z. (2004). Measurement structure of the Turkish translation of the Child Behavior Checklist using confirmatory factor analytic approaches to validation of syndromal constructs. *Journal of Abnormal Child Psychology, 32,* 335–340.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20,* 201–212.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78,* 350–365.

Goldberg, L. R., & Digman, J. M. (1994). Revealing structure in the data: Principles of exploratory factor analysis. In S. Strack & M. Maurice (Eds.), *Differentiating normal and abnormal personality* (pp. 216–242). New York: Springer.

Hartman, C. A., Hox, J., Auerbach, J., Erol, N., Fonseca, A. C., Mellenbergh, G. J., et al. (1999). Syndrome dimensions of the Child Behavior Checklist and Teacher Report Form: A critical empirical evaluation. *Journal of Child Psychology and Psychiatry, 40,* 1095–1116.

Hotelling, H. (1933). Analysis of complex statistical variables into principal components. *Journal of Educational Psychology, 24,* 417–441, 498–520.

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Ivanova, M. Y., Achenbach, T. M., Dumenci, L., Rescorla, L. A., Almqvist, F., Weintraub, S., et al. (2007). Testing the 8-syndrome structure of the Child Behavior Checklist in 30 societies. *Journal of Clinical Child and Adolescent Psychology, 36,* 405–417.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bathiche, N., et al. (2007). Testing the Teacher's Report Form syndromes in 20 societies. *School Psychology Review, 36,* 468–483.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bilenberg, N., et al. (2007). The generalizability of the Youth Self-Report syndrome structure in 23 societies. *Journal of Clinical and Consulting Psychology, 75,* 729–738.

Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 31,* 165–178.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 44–53.

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

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320–341.

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

McDonald, R. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Muraki, E., & Bock, R. D. (2003). PARSCALE 4: IRT item analysis and test scoring for rating-scale data. Chicago: Scientific Software International.

Muraki, E., & Engelhard, G. (1985). Full information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9,* 417–430.

Muthén, L. K., & Muthén, B. O. (1998–2004). *Mplus: User's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219–241). Mahwah, NJ: Erlbaum.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81,* 93–103.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8,* 164–184.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17, 34*(4, Pt. 2).

Spector, P. E. (1992). *Summated rating scales: An introduction.* Newbury Park, CA: Sage.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 312–317.

Widaman, K. F. (2004, May). *Common factors versus components: Principals and principles, errors and misconceptions.* Presented at the Factor Analysis at 100 conference, L. L. Thurstone Psychometric Lab, University of North Carolina at Chapel Hill.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software International.

Wright, B. D. (1999). Fundamental measurement of psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Erlbaum.

Yu, C. Y., & Muthén, B. O. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes* (Technical Report). Los Angeles: University of California at Los Angeles, Graduate School of Education & Information Studies.

Received March 27, 2007
Revision received November 5, 2007
Accepted November 6, 2007 ■