JOURNAL OF APPLIED MEASUREMENT, *3*(2), 205–231
Copyright© 2002

Detecting and Evaluating the Impact of Multidimensionality using Item Fit Statistics and Principal Component Analysis of Residuals

Everett V. Smith, Jr.
The University of Illinois at Chicago

The purpose of this research is twofold. First is to extend the work of Smith (1992, 1996) and Smith and Miao (1991, 1994) in comparing item fit statistics and principal component analysis as tools for assessing the unidimensionality requirement of Rasch models. Second is to demonstrate methods to explore how violations of the unidimensionality requirement influence person measurement. For the first study, rating scale data were simulated to represent varying degrees of multidimensionality and the proportion of items contributing to each component. The second study used responses to a 24 item Attention Deficit Hyperactivity Disorder scale obtained from 317 college undergraduates. The simulation study reveals both an iterative item fit approach and principal component analysis of standardized residuals are effective in detecting items simulated to contribute to multidimensionality. The methods presented in Study 2 demonstrate the potential impact of multidimensionality on norm and criterion-reference person measure interpretations. The results provide researchers with quantitative information to help assist with the qualitative judgment as to whether the impact of multidimensionality is severe enough to warrant removing items from the analysis.

Requests for reprints should be sent to Everett Smith, University of Illinois at Chicago, 1040 West Harrison Street, M/C 147, Chicago, IL, 60607, e-mail: evsmith@uic.edu.

Unidimensionality is frequently defined as a single latent trait being able to account for the performance on items. Bejar (1983, p.31) expanded this strict definition of unidimensionality and notes that 'unidimensionality does not imply that performance on the items is due to a single psychological process. In fact, a variety of psychological processes are involved in the act of responding to a set of items. However, as long as they function in unison—that is, the performance on each item is affected by the same processes and in the same form—unidimensionality will hold.'

Stout (1987) stated that there are at least three reasons why it is important that responses to an assessment represent a unidimensional construct. First, an assessment that purports to measure the level of a construct should not be significantly influenced by varying levels of one or more other abilities. That is, the measurement of any object or entity describes only one attribute of the object measured. This is a universal characteristic of all measurement (Thurstone, 1931). Second, an assessment to be used in identifying individual differences or ordering persons on some attribute must measure a unidimensional construct. This is a requirement for two individuals with the same score to be considered similar. Third, unidimensionality must hold before a total score is calculated under true score theory (TST) or ability estimated using many of the popular item response theory (IRT–2P, 3P and extensions) and Rasch models as violations of the unidimensionality requirement may bias item and person estimates (Reckase, 1979; Harrison, 1986).

Unfortunately, when analyzing real data, one way in which all unidimensional IRT and Rasch models are incorrect is in the requirement of unidimensionality; as there are always other cognitive, personality, and test-taking factors that influence performance. Therefore, unidimensionality should not be viewed as a dichotomous yes or no decision, but rather as a continuum. A relevant research question then becomes, 'At what point on the continuum does multidimensionality threaten the interpretation of the item and person estimates?' However, prior to evaluating the impact of multidimensionality, methods for assessing the degree to which a set of data represents a unidimensional construct must be employed. Regrettably, there is no agreed upon method for assessing unidimensionality.

Among the earlier methods used to assess the unidimensionality requirement was linear factor analysis of phi or tetrachoric correlations with a plotting of the eigenvalues from largest to smallest, looking for a dominant first factor and a high ratio of the first to the second eigenvalue (Hambleton and Swaminathan, 1985). Unfortunately, linear factor analysis of phi or tetrachoric correlations may lead to an overestimation of the number of factors underlying the data (Bock, Gibbons, and Muraki, 1988; Gibbons, Clark, Cavanaugh, and Davis, 1985; Hambleton and Swaminathan, 1985) and spurious factors due to the difficulty of test items (Bernstein and Teng, 1989; Mislevy, 1986; Hulin, Drasgow, and Parsons, 1983).

More recent methods for assessing the unidimensionality requirement include full information factor analysis (Bock, Gibbons, and Muraki, 1988; Muraki and Engelhard, 1985), methods based on the principle of 'essential' unidimensionality (Stout, 1987; 1990), and modified parallel analysis (Drasgow and Lissak, 1983).

Full-information factor analysis is based upon the notion that all the information available from the entire response matrix is used rather than just the covariance or correlation matrix. Full-information methods are considered analogous to fitting a two or three-parameter normal-ogive IRT model (Embretson and Reise, 2000). Modified parallel analysis compares the eigenvalues of a real data set to those from a unidimensional simulated data set based on the IRT parameters obtained from the real data. Scree plots are examined to assist in determining when violations of unidimensionality are too severe to allow satisfactory parameter estimation. Essential unidimensionality is based on the premise that a dominant dimension exists with the possible presence of several minor dimensions and that the dominant dimension is so strong that the trait estimates are not affected by the presence of the smaller dimensions. Responses are considered as satisfying essential unidimensionality when the average between-item residual covariance, after fitting a one factor model, approaches zero as the length of the test increases.

Assessing Dimensionality in Rasch Measurement

In addition to the reasons cited previously, the use of linear factor analytic models are not appropriate methods for assessing the unidimensionality requirement of Rasch models as these methods assume a normal distribution of the data, whereas Rasch models make no such assumption (Slinde and Linn, 1979). Full-information factor analytic methods also do not seem appropriate as fitting data to a 2 or 3-parameter model contradicts the requirements of Rasch models (Smith, 2001). Modified parallel analysis and methods based on 'essential' unidimensionality have not been widely adopted by Rasch practioners. This lack of adoption may be due to the historically different approaches to assess unidimensionality taken by researchers in the two measurement paradigms. Assessment of unidimensionality for IRT models appears to predominantly use methods external to the model prior to fitting the model to the data. Whereas within the context of Rasch modeling, the fit of the data to the unidimensionality requirement is often addressed internally using, for example, the discrepancy between the observed and model expected responses. A notable exception to the use of methods external to the model for assessing unidimensionality found in the IRT literature is based on the work of Bejar (1980). Bejar proposed splitting the items into meaningful groups (e.g., based on subject matter content), calibrating each set of items and the entire set of items, and then plotting the item difficulties from the meaningful item groups versus the item difficulties from the entire set of items. If the item difficulties are not equal within measurement error, then the probability of passing items at fixed levels of ability will differ. This would imply that performance on items depends on which items are included in the test, contradicting unidimensionality. Within the context of Rasch models, this method would seem to be inappropriate if the calibrations of the item subsets are based on responses from a common group of individuals. Because the item raw score is a sufficient statistic for estimation of item difficulty, identical item raw scores would be obtained regardless of the subset of items any specific item was placed in. This would result in identical item difficulty estimates, after adjusting for a shift constant, for an item calibrated in two or more different item subsets when responses from a common group of individuals are used for the calibrations.

There are two main approaches to conducting statistical tests of fit in Rasch measurement (Andrich, 1988). The first involves estimating the model parameters and then checking how well various portions of the data can be recovered. The second entails the condition that the parameter estimates should be invariant across different partitions (i.e., different groups) of the data and involves comparing parameter estimates from such partitions. This research will focus on the first approach, using item level fit statistics that compare the observed and expected responses (Smith, 1996; Wright, 1996). Readers interested in various methods to assess fit in the context of Rasch measurement should consult Andrich (1988), Anderson (1973), Glas (1988), Kelderman (1984), Glas and Verhelst (1995a, 1995b), and van den Wollenberg (1982).

The item level fit statistics used in this research are based on the chi-square fit statistics proposed by Wright, Mead, and Draba (1976) and Wright and Panchapakesan (1969) and their transformation into statistics based on the t-distribution (Wright and Stone, 1979). Given the estimated item and person parameters, for each observation X_{ni}, the derivation of item fit begins with a calculation of an expected (predicted) value:

$$E_{ni} = \sum_{k=0}^{mi} k\,(P_{nik})$$

where k is the available categories for observations and P_{nik} is the probability of person n being observed in category k on item i.

Next a score residual, Y_{ni}, and standardized residual, Z_{ni}, are calculated:

$$Y_{ni} = X_{ni} - E_{ni}$$

$$Z_{ni} = Y_{ni} / (W_{ni})^{1/2}$$

where $W_{ni} = \sum_{k=0}^{mi} (k-E_{ni})^2 P_{nik}$ is the variance of $X_{ni.}$

These standardized residuals are squared and summed to form a chi-square statistic,

$$\chi^2 = \sum_{n=1}^{N} Z_{ni}^2$$

which, when divided by N, yields the Mean-square Outfit statistic:

$$\text{Mean-square Outfit } (u_i) = (\sum_{n=1}^{N} Z_{ni}^2) / N$$

which has a variance of

$$s_i^2 = \sum_{n=1}^{N} (C_{ni} / W_{ni}^2) \; / N^2 - 1/N$$

where $C_{ni} = \sum_{k=0}^{mi} (k-E_{ni})^4 P_{nik}$ is the kurtosis of $X_{ni.}$

An inspection of s_i^2 will show that the standard deviation varies depending on the number of persons and W_{ni} and so varies from item to item and sample to sample. Since this produces difficulties in calculating a generalizable cutoff for determining adequate data-model fit, the mean-square statistic is transformed into a t-statistic with an approximate unit normal distribution:

$t_i = (u_i^{1/3} - 1)(3 / s_i) + (s_i / 3)$.

These Outfit statistics (mean-square and standardized) are sensitive to a few off-targeted (i.e., a large logit difference between an item's difficulty and a person's ability) improbable responses. To diminish this effect on the Outfit statistic, the Infit statistic weighs the squared standardized residual, Z_{ni}^2, by their individual variances, W_{ni}, (larger variances exist for targeted observations, smaller variances for off-targeted observations). Mean-square Infit is calculated as:

$$\text{Mean-square Infit } (v_i) = \sum_{n=1}^{N} W_{ni} Z_{ni}^2 / \sum_{n=1}^{N} W_{ni}$$

$$= \sum_{n=1}^{N} Y_{ni}^2 / \sum_{n=1}^{N} W_{ni}$$

which has a variance of

$$q_i^2 = \sum_{n=1}^{N} (C_{ni} - W_{ni}^2) / (\sum_{n=1}^{N} W_{ni})^2$$

which again may be transformed into a t-statistic with an approximate unit normal distribution:

$t_i = (v_i^{1/3} - 1)(3 / q_i) + (q_i / 3)$.

The t-statistic, also referred to as the standardized fit statistic, has an approximate 5% Type I error rate using a ± 2 criterion (Smith, 1991). Values greater than +2 indicate departures from unidimensionality; values less than -2 indicate overly predictable response patterns, possibly due to violations of the local independence requirement.

Smith (1992, 1996) and Smith and Miao (1991, 1994) compared the utility of the standardized item outfit statistic using a greater than +2 criterion against principal component analysis (PCA) of raw data without rotation for dichotomous and polytomous data. The simulated data varied the level of covariation between two components and the number of items contributing to each component. The conclusions were simple. When the components had approximately equal number of items contributing to each component and the components were not highly correlated, PCA was better able to detect multidimensionality. When the two components were highly correlated or the majority of items contributed to one component, the standardized item outfit statistic was better able to detect departures from unidimensionality. In summarizing this research, Smith (1996) concluded that when the intention is to create a unidimensional variable, few items would be expected to contribute to the second component and the correlation between the components should be high. It is under these conditions in which the Rasch item fit approach detects departures from unidimensionality more accurately than PCA without rotation. A potential limitation of these studies is the use of observed ordinal data with a parametric technique (PCA) that assumes at least an interval scale of measurement.

The purpose of this research is twofold. First is to extend the work of Smith (1992, 1996) and Smith and Miao (1991, 1994) to involve a comparison of item fit statistics with PCA of standardized residuals as tools for assessing the unidimensionality requirement of Rasch models. Second is to demonstrate methods for determining the impact of multidimensionality on person measurement.

Method

Study 1: Assessing Unidimensionality

Simulation of Rating Scale Data. The data for Study 1 were simulated to represent varying degrees of common variance between two components and varying proportion of items representing each component. The generated levels of common variance ranged from .00 to .90 in increments of .10, thus providing for 10 levels of common variance. The number of items representing each component also varied across three different ratios: 25:5, 20:10, and 15:15. A total of 30 items were chosen to represent an 'average' length for a rating scale. For each data set a total of 500 simulated persons were used to represent an 'average' number of participants.

For each person two sets (X_i and Y_i) of independent unit normal ability distributions were generated. From the two distributions the correlated ability distribution (Y_c) was created by substituting one of the common variance values in the following equation:

$$Y_c = (a*X_i + Y_i*SQRT(1\text{-}a**2))*\ S_c + M_c$$

where X_i and Y_i represent the two sets of independent unit normal ability distributions, Y_c represents the ability distribution correlated with X_i at a specified level of a. S_c and M_c represent the standard deviation and mean of the Y_c distribution, respectively. S_c and M_c were specified to be 1 and 0, respectively, for all simulated data. The results of the simulation of the correlated ability distribution are shown in Table 1.

Table 1

Results for the simulation of the correlated ability distribution (Y_c)

Level of Common Variance	Specified Correlation	Observed Correlation	Observed Mean (SD)	Observed Skewness	Observed Kurtosis
.00	.00	-.04	.04 (1.00)	.07	.31
.10	.32	.33	.01 (1.04)	.02	.03
.20	.45	.46	.01 (1.02)	.06	-.02
.30	.55	.56	-.03 (0.97)	.13	.17
.40	.63	.63	-.04 (1.00)	.09	-.05
.50	.71	.71	.00 (1.01)	.04	-.20
.60	.77	.78	-.03 (1.04)	-.25	.18
.70	.84	.84	-.02 (1.03)	.07	-.39
.80	.89	.89	.03 (1.00)	-.06	.13
.90	.94	.95	.07 (1.05)	.02	-.19

For each level of common variance, the two abilities (X_i and Y_c) for each person were then used to create simulated responses. The X_i ability was used to generate the responses to the items representing the X component and the Y_c values were used to generate the responses to the items representing the Y component using the Rasch rating scale model as implemented by WINSTEPS (Linacre, 2001). The item difficulties used in the simulations were uniformly distributed in sets of five items (-1, -.5, 0, .5, and 1) so the number of items representing each component would not have an influence on the mean or distribution of the item difficulties for that data set (Smith, 1996). The step values were fixed at -1, -.33, .33, and 1, thus representing a 5-point rating scale.

Analysis.

All simulated data sets were analyzed using WINSTEPS (Linacre, 2001). Linacre (1992) suggested a three step process to investigate dimensionality: addressing negative corrected item-total correlations for obvious off-dimension behavior or a mistake in coding the data (e.g., forgetting to reverse code items), diagnosing idiosyncratic response patterns using fit statistics, and finally looking for patterns among the standardized residuals using PCA without rotation. Only the item fit statistics and PCA of the standardized residuals were used to investigate multidimensionality as the corrected item-total correlations were unlikely to provide useful information given the data were not simulated to represent negatively correlated dimensions. For this research the standardized Infit and Outfit statistics were used rather than the mean-square fit statistics due to a preference for fit statistics with approximate unit normal sampling distributions (Smith, 2000).

Why PCA of standardized residuals?

The Rasch model constructs a one dimensional measurement system regardless of the dimensionality of those data (Linacre, 1998a). The ideal would be that all information in the data be explained by the latent variable. Then, the residuals would represent random noise, which when standardized would follow a normal distribution. Furthermore, the residuals would be independent of each other (i.e., each residual represents a unique factor). As a consequence, all elements of an inter-item residual correlation matrix would be zero if the data fit the model.

However, each observation, will to some degree, contain its own characteristic features. These characteristic features are manifested as the difference between what the Rasch model predicts and what is observed (i.e., the residual). Therefore, to test the hypothesis that the data fit the model, WINSTEPS (Linacre, 2001) asserts that all the residual variance is due to common factors and places 1's in the diagonal of the inter-item residual correlation matrix and the empirical correlations among the standardized residuals in the off-diagonal elements. Principal components analysis of these standardized residuals identifies characteristics shared in common among items. These are often indications of secondary structures of subdimensions within the data (Linacre, 1998b) that may warrant action and diagnosis (Study 2).

Study 2: Evaluating the impact of multidimensionality on person measurement

Sample. Data come from Smith and Johnson (2000). Three hundred seventeen (90 men, 225 women) undergraduate students (87 freshman, 84 sophomores, 95 juniors, and 51 seniors) drawn from midwestern and northwestern universities provided responses to the Adult Behavior Checklist – Revised (ABC-R). The average age was 19.69 years (SD=1.09). The average self-reported grade point average was 3.25 (SD=.51). The ethnic background of participants included 4 African-Americans, 4 Hispanic/Latinos, 12 Native Americans, 285 Caucasians, 6 Asians, and 4 with unidentified 'other' ethnic backgrounds.

Instrumentation. The ABC-R (see Appendix; Smith and Johnson, 2000) is a 24 item self-report instrument designed to screen for Attention Deficit Hyperactivity Disorder (ADHD) symptomatology in college students. Items one through seventeen are designed to measure the Attentive aspect of ADHD; items eighteen through twenty-four Hyperactivity/Impulsivity. The response format for all items consists of a 4-point Likert-type scale ranging from 1 to 4 with labels of "Never/Not At All", "Sometimes/Just A Little", "Often/Pretty Much", and "Very Often/Very Much".

Analysis. Data from all 24 items were analyzed using the Rasch rating scale model (Andrich, 1978) as implemented by WINSTEPS (Linacre, 2001). Corrected item-total correlations, standardized Infit and Outfit statistics, and PCA of standardized residuals without rotation were used to detect departures from the unidimensionality requirement. Given that ADHD is considered a multidimensional construct, it is expected that these analyses will reveal more than one dimension underlying the responses. However, multidimensionality only becomes a problem when data represent two or more dimensions so disparate or distinct that it is no longer clear what dimension the Rasch model is defining (lacks construct validity) or when the different subsets of items would lead to different norm (NR) or criterion-reference (CR) decisions. Therefore, after the items contributing to multidimensionality were identified, the impact of multidimensionality on person measure interpretations was investigated.

Evaluating impact on person measure interpretation. The first step was to conduct a separate calibration of the items identified as contributing to multidimensionality and then account for the change in the local origin of the two analyses (Wright and Stone, 1979). For the I items identified as contributing to multidimensionality, the simplest method to account for the change in the local origin is to sum the differences of the I item difficulty estimates from the two analyses and calculate an average:

$$\Sigma (d_i - d'_i) / I$$

where d_i is the item difficulty estimate of item i when analyzed with all 24 items, d'_i is the item difficulty estimate of item i when calibrated separately with the other items contributing to multidimensionality, and I is the number of items identified as contributing to multidimensionality. This average is often called a shift constant and represents the change in the local origin for the two analyses.

The next step was to adjust the person measures obtained from the calibration of the I items identified as contributing to multidimensionality for the shift constant. Adding the shift constant to the person measures from the separate calibration brings the separate calibration person measures onto the same scale as person measures from the calibration of all 24 items (Wright and Stone, 1979). An alternative to comparing the person measures from the calibration of the I items contributing to multidimensionality to the person measures from the calibration of all 24 items would be to compare the person measures from a calibration of the I items contributing to multidimensionality to the person measures from a calibration of the remaining (24 – I) items. In this scenario, two shift constants would need to be computed to bring the person measures from the separate calibration of the I and (24 – I) items onto the same scale.

To see if different NR interpretations would be made based on the person measures from each analysis, the pairs of person measures were correlated. Given high person reliabilities from each analysis, a correlation substantially less than one would indicate that the rank order of the individuals differs depending on which items are included in the calibration, indicating the items measure different constructs.

CR interpretations were evaluated with two methods. The first involved conducting a series of independent t-tests (or equivalently constructing 95% control lines in a plot of the person measures) with the pairs of person measures and the associated standard errors using

$$t = (b_s - b_c) / SQRT\ (SE_{bs}^2 + SE_{bc}^2)$$

where b_s is the person measure adjusted for the shift in the local origin, b_c is the person measure from the analysis of all 24 items, and SE_{bs} and SE_{bc} are the respective standard errors of the person measures. A statistically significant t-test would indicate the level of the trait differs depending on which items are included in the calibration. This would indicate multidimensionality, as under Rasch model conditions, if the data fit the model then analysis of any subset of items should produce equivalent person measures, within measurement error. The second method employed required the existence of a cutscore. Determining the number of decision changes (e.g., pass to fail and fail to pass) based on the two calibrations provided another indicator of the impact of multidimensionality on person measure interpretation.

Results

Study 1: Assessing Unidimensionality

Interpretation of the PCA depends in part on the choice of the critical value of the eigenvalue. To determine the best value to use to assess whether a second component exists, three independent sets of data were simulated (see Schumacker, Smith, and Bush, 2000 for a discussion of the effects of the number of replications on the stability of simulation results) using the item ratio of 30:0. The average eigenvalue for the first component of a PCA of the standardized residuals for the three simulations was 1.5, which represents 5% (1.5 / 30) of the random variation in the standardized residuals. Thus, any eigenvalue greater than 1.5 will be considered as representing the existence of a second component. Given the existence of a second component, the success of the PCA in identifying which items contribute to a second component can be evaluated at different levels. The first definition of success would require the structure coefficients (i.e., loadings) for each item set (i.e., the x items contributing to the X component and the y items contributing to the Y component) be mutually exclusive, that is the structure coefficients for each item set must not overlap. This definition could pose difficulty in terms of interpretation. This could happen since most applications of PCA only interpret item content of structure coefficients that share the same sign and are 'salient'. The requirement of mutually exclusive structure coefficients for each item set may not necessarily reflect sets of all positive versus all negative structure coefficients or structure coefficients that are deemed salient for interpretation. Therefore, the second definition of success examined the sign of the structure coefficients and the magnitude the structure coefficients would need to exceed to be considered salient for interpretation. The difficulty resides in the choice of a cutoff for a salient structure coefficient to be used in determining successful identification by the PCA of the x and y items to the X and Y components. The most objective way to determine this cutoff would be to use simulated data. Therefore, the structure coefficients from the three simulated data sets used to determine the eigenvalue cutoff were pooled together and a value was then determined that would be exceeded only 5% of the time in the data simulated to fit the model. This value was determined to be a structure coefficient of ± .38. Using this cutoff, the percentage of x and y items being correctly classified ($\geq |.38|$ on the appropriate component), incorrectly classified ($\geq |.38|$ on the incorrect component), or not classified (structure coefficient between ± .38) was determined for each level of common variance and item ratio combination. Since the y items were simulated to represent the items contributing to multidimensionality, the desired results would have the y items correctly classified on the Y component and the x items not incorrectly classified as belonging to the Y component.

Interpretation of the standardized Infit and Outfit statistics was accomplished by comparing the value of the fit statistics to a critical value of 2. The presence of a second component was determined by evaluating the percentage of X and Y component items with fit statistic values greater than 2. The ideal situation would be all fit statistics for the X component items be within the Type I error rate and all Y component items have fit statistics greater than 2. This would indicate all items contributing to the correlated component Y being identified as misfitting (i.e., contributing to a second component) and none of the items contributing to the primary component X being identified as contributing to a second component. Table 2 shows the percentage of misfitting x and y items, eigenvalues for the residual principal component, and the range of x and y structure coefficients for each item ratio and level of common variance. Table 3 shows the percentage of x and y items correctly classified, incorrectly classified, or not classified using the ± .38 cutoff for a salient structure coefficient.

Using a cutoff of 2, the standardized Infit and Outfit statistics displayed in Table 2 were able to correctly identify all the items contributing to the second component Y and not falsely identify any item that was specified to belong to the first component X for the levels of common variance up to .60 for the 25:5 item ratio and for the 0 and .10 levels of common variance for the 20:10 item ratio. Moderate success, defined as greater than 50% correct identification of the y items by at least one of the fit statistics and less than or equal to a 5% (the Type I error rate) false identification of the x items, was found for the .70 level of common variance for the 25:5 item ratio and for the .20 to .50 levels of common variance for the 20:10 item ratio. The standardized fit statistics performed poorly for the remaining levels of common variance and item ratio combinations with the ability of the fit statistics to identify the y items deteriorating as the level of common variance increased and as the ratio of the number of x and y items approached one.

For all levels of common variance and all item ratios, the eigenvalue for the PCA of the standardized residuals exceeded the cutoff of 1.5, indicating the existence of a second component under all conditions included in this study. The range of the structure coefficients for the x and y items were mutually exclusive for all conditions except the .90 level of common variance for the 20:10 item ratio and the .70, .80, and .90 levels of common variance for the 15:15 item ratio. For the conditions in which the range of structure coefficients were mutually exclusive, the PCA of the standardized residuals identified the x items as contributing to the X component (negative structure coefficients) and y items as contributing to the Y component (positive structure coefficients) with the exception of the .50 to .90 levels of common variance for the 25:5 item ratio.

Table 2

Percentage of misfitting x and y items, eigenvalues for the residual principal component, and the range of x and y item structure coefficients for each x:y item ratio and level of common variance.

Item Ratio for X and Y Components		Level of Common Variance Shared Between X_j and Y_c									
		0	10	20	30	40	50	60	70	80	90
25:5	% x with Infit > 2	0	0	0	0	0	0	0	0	0	0
	% x with Outfit > 2	0	0	0	0	0	0	0	4	0	0
	% y with Infit > 2	100	100	100	100	100	100	100	60	20	20
	% y with Outfit > 2	100	100	100	100	100	100	100	60	40	20
	Eigenvalue	4.5	3.7	3.4	2.8	2.8	2.5	2.2	1.9	1.7	1.7
	Range of x coefficients	-.35,-.16	-.28,-.09	-.28,-.09	-.26,-.07	-.27,-.02	-.27,.01	-.29,.01	-.27,.07	-.30,.16	-.27,.25
	Range of y coefficients	.73,.82	.67,.78	.68,.74	.59,.70	.60,.69	.58,.67	.50,.63	.48,.59	.39,.53	.33,.52
20:10	% x with Infit > 2	0	0	0	0	0	0	5	0	5	5
	% x with Outfit > 2	0	0	0	0	0	0	5	0	0	5
	% y with Infit > 2	100	100	100	60	70	60	20	0	10	10
	% y with Outfit > 2	100	100	90	70	60	40	20	30	20	10
	Eigenvalue	8.3	6.5	5.5	4.6	4.4	3.9	3.4	2.9	2.3	1.7
	Range of x coefficients	-.50,-.33	-.46,-.25	-.41,-.21	-.42,-.19	-.39,-.18	-.39,-.18	-.35,-.15	-.35,-.08	-.35,-.05	-.32,.03
	Range of y coefficients	.64,.72	.55,.67	.51,.63	.46,.59	.45,.55	.42,.56	.36,.52	.34,.48	.24,.48	-.25,.60
15:15	% x with Infit > 2	6.7	0	6.7	13.3	0	0	6.7	0	0	6.7
	% x with Outfit > 2	0	0	0	6.7	0	6.7	0	0	0	0
	% y with Infit > 2	0	0	6.7	6.7	13.3	0	0	0	0	6.7
	% y with Outfit > 2	0	0	6.7	6.7	6.7	0	0	0	0	6.7
	Eigenvalue	9.4	6.7	6.0	4.7	4.4	3.6	3.0	2.5	2.5	2.5
	Range of x coefficients	-.60,-.48	-.53,-.38	-.50,-.35	-.44,-.30	-.47,-.27	-.47,-.20	-.46,-.11	-.19,.57	-.42,.51	-.46,.46
	Range of y coefficients	.51,.61	.36,.55	.34,.50	.29,.45	.28,.47	.21,.47	.12,.46	-.44,.27	-.45,.49	-.47,.50

Table 3

Percentage of x and y itemsthat are correctly classified, incorrectly classified, or not classified by the PCA of standardized residuals for each x:y item ratio and level of common variance.

		Level of Common Variance Shared Between X_j and Y_c									
Item Ratio for X and Y Components		0	10	20	30	40	50	60	70	80	90
25:5	% x correct	0	0	0	0	0	0	0	0	0	0
	% x incorrect	0	0	0	0	0	0	0	0	0	0
	% x not classified	100	100	100	100	100	100	100	100	100	100
	% y correct	100	100	100	100	100	100	100	100	100	80
	% y incorrect	0	0	0	0	0	0	0	0	0	0
	% y not classified	0	0	0	0	0	0	0	0	0	20
20:10	% x correct	90	40	30	10	10	5	0	0	0	0
	% x incorrect	0	0	0	0	0	0	0	0	0	0
	% x not classified	10	60	70	90	90	95	100	100	100	100
	% y correct	100	100	100	100	100	100	80	80	60	40
	% y incorrect	0	0	0	0	0	0	0	0	0	0
	% y not classified	0	0	0	0	0	0	20	20	40	60
15:15	% x correct	100	100	93.33	66.67	60	33.33	26.67	20	13.33	6.67
	% x incorrect	0	0	0	0	0	0	0	0	13.33	6.67
	% x not classified	0	0	6.67	33.33	40	66.67	73.33	80	73.33	86.67
	% y correct	100	86.67	86.67	66.67	60	33.33	33.33	26.67	13.33	6.67
	% y incorrect	0	0	0	0	0	0	0	0	13.33	6.67
	% y not classified	0	13.33	13.33	33.33	40	66.67	66.67	73.33	73.33	86.67

The classification results in Table 3 complemented the results of the fit statistics and extended the range of conditions under which multidimensionality was detected. For the conditions under which the fit statistics performed with at least moderate success (i.e., for the 0 to .70 levels of common variance for the 25:5 item ratio and for the 0 to .50 levels of common variance for the 20:10 item ratio), the classification results were able to correctly identify 100% of the y items as belonging to the Y component and not incorrectly classify any x items as contributing to the Y component (0%). Using a minimum of a 50% correct classification rate of the y items and a 5% rate of incorrect classification of the x items to the Y component, the classification results extended the range of conditions under which multidimensionality was detected to include the .80 and .90 levels of common variance for the 25:5 item ratio, the .60 to .80 levels of common variance for the 20:10 item ratio, and the 0 to .40 levels of common variance for the 15:15 item ratio.

Table 4

Item fit information for all 24 items

Item Number	Measure	Error	Infit Zstd	Outfit Zstd	Score Corr.	Item Names - Dimension
20	-1.11	.08	7.1	7.2	.30	DRIVENBYMOTOR-HYP
19	-1.62	.08	5.1	6.1	.34	ONTHEGO-HYP
21	-.61	.08	4.1	4.0	.40	TALKEXCESS-HYP
23	.21	.09	3.5	3.1	.54	WAITINGTURN-HYP
18	.06	.09	3.3	2.8	.54	REMAINQUIET-HYP
15	1.17	.11	2.8	1.2	.52	LOSETHINGS-ATTN
22	.40	.10	1.6	1.9	.40	BLURTANSW-HYP
6	1.24	.11	1.3	.9	.41	NOLISTEN-ATTN
24	.12	.09	.8	.3	.49	INTERRUPTOTH-HYP
8	.49	.10	.8	.4	.49	FAILFINSCH-ATTN
9	.10	.09	.1	-.2	.55	FAILFINCHO-ATTN
5	-.56	.08	-.4	-.6	.55	SUSTATTN-ATTN
11	.65	.10	-.7	-1.2	.48	DIFFORG-ATTN
13	-.65	.08	-1.4	-1.0	.48	DISLIKETASKS-ATTN
10	1.57	.12	-1.1	-1.4	.45	FAILFINWRK-ATTN
16	-1.15	.08	-1.8	-2.1	.45	EASYDISTRAC-ATTN
17	.07	.09	-1.9	-2.2	.55	FORGETFUL-ATTN
12	-.07	.09	-2.0	-2.3	.55	AVOIDTASKS-ATTN
2	.18	.09	-2.4	-2.1	.29	ATTATWRK-ATTN
7	1.10	.11	-2.3	-2.7	.54	FOLLINSTR-ATTN
14	-.45	.09	-3.3	-3.3	.57	RELUCENGAGE-ATTN
4	.24	.09	-3.9	-3.9	.35	MISATWRK-ATTN
1	-.84	.08	-6.7	-6.7	.48	ATTATSCH-ATTN
3	-.55	.08	-7.0	-6.9	.48	MISATSCH-ATTN

Note: HYP = Hyperactive/Impulsive, ATTN = Attentive.

Study 2: Evaluating the impact of multidimensionality on person measurement

Table 4 displays the corrected item-total correlations (Score Corr.) and fit statistics for the calibration of all 24 ABC-R items presented in fit order. Table 4 also contains the item location in the ABC-R (Item Number), the estimated item endorsability (Measure) and associated standard error (Error).

None of the corrected item-total correlations were negative or close to zero, indicating no obvious off-dimension responses or miscoded data. Five of the seven (71.43%) items hypothesized to define the Hyperactive/Impulsive aspect of ADHD had fit statistics greater than 2; one item (5.88%) hypothesized to define the Attentive aspect of ADHD had a fit statistic greater than 2. In order to compare these percentages to the results from Study 1, an estimate of the magnitude of the relationship between the Hyperactive/Impulsive and Attentive aspects of ADHD was required. Previous research estimated this correlation to be around .40 (Smith and Johnson, 1998). Using the closest item ratio of 20:10 from the simulation in Study 1 that approximates the 17:7 item ratio of the ADHD items, the results from the fit analysis of the ADHD items closely approximated the percentages in Table 2.

Table 5 and Figure 1 display the results of the PCA of the standardized residuals. The residual component had an eigenvalue of 3.9, representing 16.25% (3.9 / 24) of the residual variance. Based on the previous simulations, it appears that this residual component is explaining more than just random variation. It is evident from Figure 1 that the 7 Hyperactive/Impulsive items (HYP) are clearly separated from the 17 Attentive items (ATTN). Noteworthy is the fact that the majority (6 of 7) of HYP items (analogous to the y items in the simulations in Study 1 since the HYP items represent the minority of items on the ABC-R) have what may be considered salient structure coefficients while the ATTN items (analogous to the x items in Study 1) have a mixture of salient and non-salient structure coefficients. The combined evidence from the fit analysis and the PCA of the standardized residuals has identified all items that are hypothesized to measure Hyperactivity/Impulsivity as contributing to a second dimension.

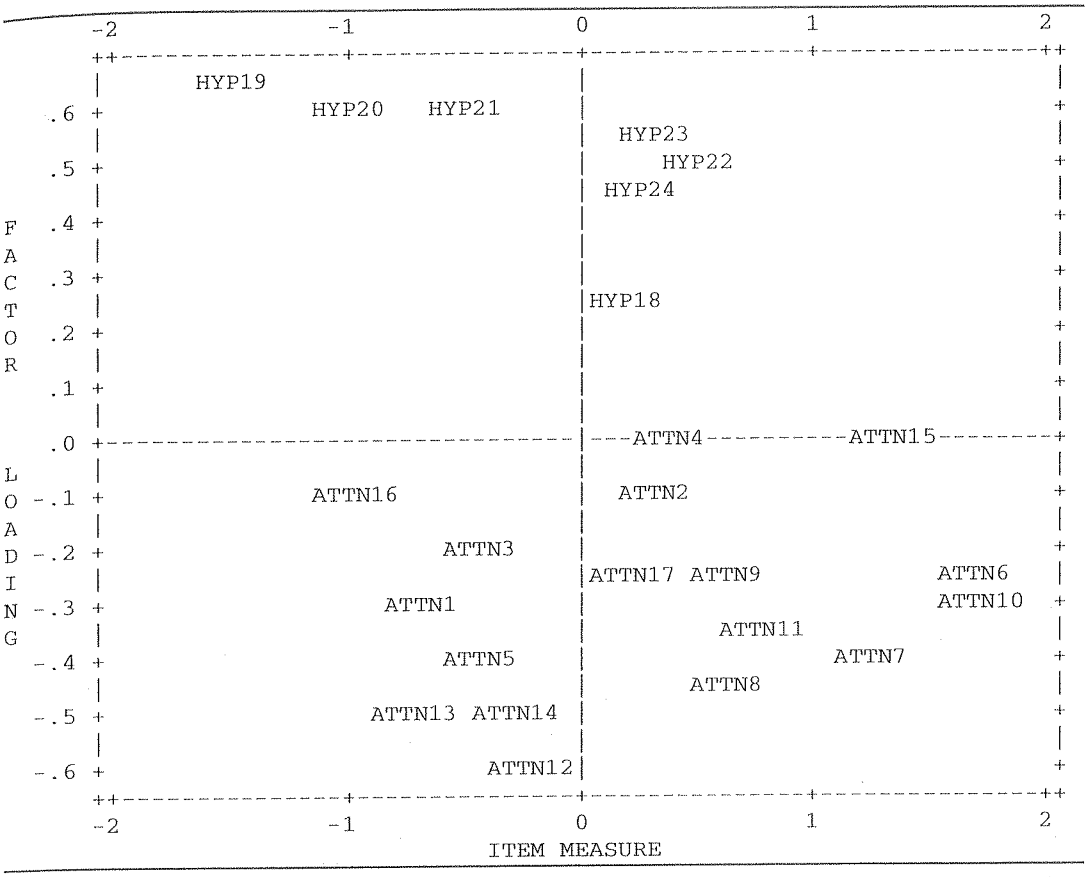
Impact on person measure interpretation. Table 6 demonstrates the process of calculating the shift constant. The sum of the item shift values was -2.52. When divided by the number of items (7), this yields a shift constant of -.36.

Table 5

Structure coefficients from the PCA of the standardized residuals for all 24 items

Loading	Item Number	Item Names
.64	19	You act as if you are "on the go"
.62	21	You talk excessively
.62	20	You act as if "driven by a motor"
.53	23	You have difficulty awaiting your turn
.50	22	You blurt out answers before questions have been completed
.45	24	You interrupt on others (e.g. butt into conversations or activities)
.24	18	You have difficulty remaining quiet during leisure activities
.01	4	You make careless mistakes at work
-.58	12	You avoid tasks that require sustained mental effort (e.g. homework, schoolwork)
-.51	14	You are reluctant to engage in tasks that require sustained mental effort (e.g. homework, schoolwork)
-.51	13	You dislike tasks that require sustained mental effort (e.g. homework, schoolwork)
-.45	8	You fail to finish schoolwork
-.41	5	You have difficulty sustaining your attention to tasks/activities
-.40	7	You do not follow through on instructions
-.36	11	You have difficulty organizing tasks/activities
-.31	1	You fail to pay close attention to details in school
-.28	10	You fail to finish work duties
-.26	17	You are forgetful in daily activities
-.25	9	You fail to finish chores
-.21	6	You do not listen when directly spoken to
-.20	3	You make careless mistakes in school
-.11	2	You fail to pay close attention to details at work
-.08	16	You are easily distracted by extraneous stimuli (e.g. traffic noises, conversations, looking out the window)
-.02	15	You lose things necessary for tasks/activities (e.g. books, keys, tools, school assignments)

The shift constant of -.36 was added to the person measures obtained from the calibration of the 7 Hyperactive/Impulsive items in order to bring these person measures onto the same scale as the person measures from the calibration of all 24 items. Figure 2 is a plot of the person measures. The correlation displayed in this plot is .64 (.84 when corrected for attenuation using the person reliabilities of .81 and .72 from the calibration of the 24 and



```
        -2              -1               0               1               2
        ++---------------+---------------+---------------+---------------++
        |      HYP19                     |                                |
     .6 +               HYP20   HYP21    |                                +
        |                                |  HYP23                         |
     .5 +                                |     HYP22                      +
        |                                | HYP24                          |
F    .4 +                                |                                +
A       |                                |                                |
C    .3 +                                |                                +
T       |                                |HYP18                           |
O    .2 +                                |                                +
R       |                                |                                |
     .1 +                                |                                +
        |                                |                                |
     .0 +--------------------------------|---ATTN4----------ATTN15--------+
L       |                                |                                |
O   -.1 +               ATTN16           |  ATTN2                         +
A       |                                |                                |
D   -.2 +                        ATTN3   |                                +
I       |                                |ATTN17 ATTN9           ATTN6    |
N   -.3 +                    ATTN1       |                       ATTN10   +
G       |                                |        ATTN11                  |
    -.4 +                        ATTN5   |               ATTN7            +
        |                                |      ATTN8                     |
    -.5 +                   ATTN13 ATTN14|                                +
        |                                |                                |
    -.6 +                          ATTN12|                                +
        ++---------------+---------------+---------------+---------------++
        -2              -1               0               1               2
                                    ITEM MEASURE
```

Note: Component 1 explains 3.9 of 24 residual variance units. Numbers following ATTN and HYP represent the Item Number.
HYP = Hyperactive/Impulsive, ATTN = Attentive.

Figure 1. Principal component analysis of the standardized residuals for all 24 items.

7 items, respectively). This correlation indicates the rank order of individuals change (a NR interpretation) depending on the set of items analyzed. The change in rank order may be the result of at least two factors. First, the changes could be due to errors of measurement given the lack of perfect reliability (i.e., .81 and .72) from the two calibrations. Second, the changes may be due to the presence of multidimensionality, as changes in rank order should not occur if the items define the same unidimensional variable. With respect to CR interpretations, a series of independent t-tests demonstrated 49 of 317 (15.45%) of the t-tests were statistically significant. This implies the level of the trait changes depending on the set of items calibrated. This is an indication of multidimensionality because if the data fit the model, then analysis of any subset of items should produce equivalent person measures, within measurement error. If a cutscore is available, determining the number of decision changes (e.g., pass to fail and fail to pass) based on the two

Table 6

Calculating the shift constant

Item Number	Item Names	d_c	d_{sub}	Shift
18	REMAINQUIET	.06	.46	-.40
19	ONTHEGO	-1.62	-1.39	-.23
20	DRIVENBYMOTOR	-1.11	-.78	-.33
21	TALKEXCESS	-.61	-.25	-.36
22	BLURTANSW	.40	.81	-.41
23	WAITINGTURN	.21	.61	-.40
24	INTERRUPTOTH	.12	.51	-.39

Note: d_c and d_{sub} represent the item difficulty estimates from the combined (all 24 items) and separate (7 items) calibrations, respectively. Shift is the difference between d_c and d_{sub}.

calibrations will provide another indicator of the impact of multidimensionality on person measure interpretation.

In Figure 2, a hypothetical cutscore for further diagnostic evaluation at zero logits has been imposed. Participants in quadrants I and III would have the same decisions (no further evaluation or further evaluation) made regarding their level of ADHD regardless of the item set employed. The three participants in quadrant II would be sent for further evaluation using the measures obtained from the calibration of all 24 items but not sent for further evaluation if decisions were based on the measures obtained from the calibration of the 7 Hyperactive/Impulsive items. The opposite decisions would be applied to the 27 participants in quadrant IV. The consequences of these decisions would need to be considered prior to determining

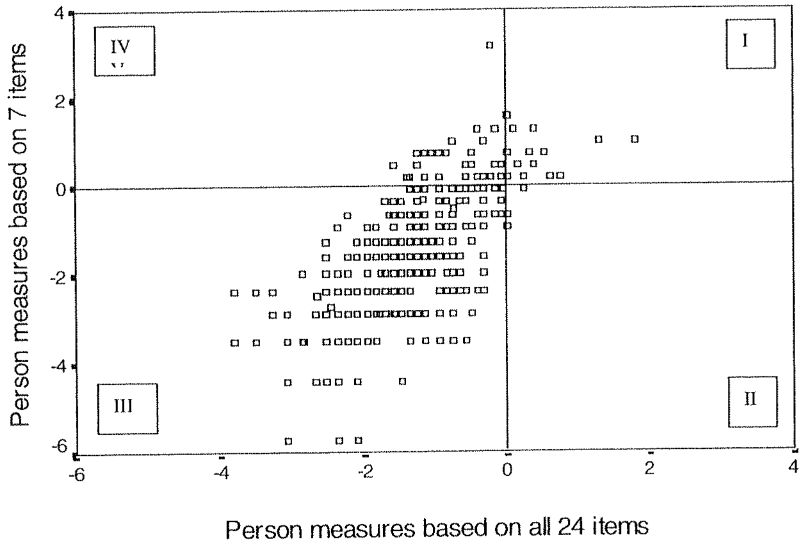


Figure 2. Plot of person measures based on 24 and 7 ADHD items.

if the items contributing to multidimensionality are adversely impacting the decisions being made. For example, if one assumes the 24 items represent the 'gold standard', the consequence of using the 7 Hyperactive/Impulsive items for decision making purposes for the 3 participants in quadrant II would be not to refer for further evaluation and potentially deny these participants treatment for ADHD. A potential consequence of referral for the participants in quadrant IV would be the costs associated with additional diagnostic procedures that are not really warranted.

Based on the cumulative evidence from the evaluation of item fit, the PCA of the standardized residuals, and the impact on person measure interpretation, the 24 items on the ABC-R appear to be measuring two aspects of ADHD. Therefore, the final decision would be to treat the Attention and Hyperactive/Impulsive items as separate variables, calibrate the two item sets separately, and report two measures for each person.

Discussion

Study 1

Study 1 compared the use of Rasch standardized fit statistics and PCA of standardized residuals for detecting departures from the unidimensionality requirement of Rasch models. The fit statistics performed well up to a level of common variance of .70 for the 25:5 item ratio and for the 20:10 item ratio up to a level of common variance of .50. As the level of common variance increased and as the ratio of the number of x to y items approached one, the ability of the fit statistics to identify the y items simulated to contribute to a second dimension deteriorated. There was no discernable advantage to relying on either Infit or Outfit. The classification results based on the PCA extended the range under which multidimensionality was detected to include the .80 and .90 levels of common variance for the 25:5 item ratio, the .60 to .80 levels of common variance for the 20:10 item ratio, and the 0 to .40 levels of common variance for the 15:15 item ratio. Again, as the level of common variance increased, the classification of the y items on the Y component declined, especially for the 15:15 item ratio.

The true performance of the fit statistics is likely to be underestimated by the results presented in Table 2. Table 2 simply presents a snapshot of the ability of the fit statistics to detect departures from unidimensionality given the current configuration of the data. Since evaluation of fit is an iterative process, once misfitting items are identified and removed, recalibration of the remaining data would likely reveal other items that may contribute to multidimensionality. For example, after removing the 7 items identified in the 20:10 item ratio for the .30 level of common variance and recalibrating the data, the remaining 3 items simulated to represent multidimensionality were subsequently identified as misfitting. Removal of these 3 items and recalibration of the remaining 20 items that were simulated to represent only one dimension did not yield any misfitting items. Thus, if this approach were actually implemented, the success rate for this ratio of items and level of common variance would be 100%.

It appears an iterative approach to fit analysis in conjunction with PCA of standardized residuals can provide insights into which items may be contributing to multidimensionality. The usefulness of these methods appears to be optimal under conditions in which the goal of the assessment design is to produce a unidimensional assessment (or several unidimensional subscales to be analyzed separately). It is under these design specifications that one would expect the majority of items to define a unidimensional variable with only a few items unexpectedly contributing to multidimensionality.

Future research should vary the factors fixed in the current investigation. Experimenting with the distribution of item difficulties, the number of items and persons, the spacing of the step calibrations, and the mean, standard deviation, skewness and kurtosis of the ability distributions will help clarify the extent to which fit statistics and PCA of standardized residuals are useful for detecting departures from unidimensionality.

Study 2

Research on applying univariate item response models to multidimensional data is popular (Asley and Forsyth, 1985; De Ayala, 1994; Drasgow and Parsons, 1983; Harrison, 1986; Reckase, 1979). However, these studies have yielded inconsistent findings (Hsu and Yu, 1989) and the impact on parameter estimation for small departures from unidimensionality remains undemonstrated (Embretson and Reise, 2000). A number of factors may contribute to the inconsistent findings. For example, different studies use different software, many of which use different estimation methods. Other potential causes for the inconsistencies may be the method used to generate the data (i.e., using a factor-analytic method or an item response model) and how researchers have defined unidimensionality. The purpose of Study 2 was not to attempt a resolution of the apparent inconsistent findings, but to demonstrate additional tools that may be employed to investigate how departures from unidimensionality impact person measure interpretation. The cumulative evidence can then be evaluated in the context and consequences of testing (Messick, 1989, 1995) to reach a decision regarding whether to discard the items identified as contributing to multidimensionality.

References

Anderson, E. B. (1973). A goodness-of-fit test for the Rasch model. *Psychometrika, 38*, 123-140.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Andrich, D. (1988). *Rasch models for measurement*. Sage University paper series on Quantitative Applications in the Social Sciences, 68. Beverly Hills: Sage Publications.

Ansley, T. M., and Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 39-48.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17*, 283-296.

Bejar, I. I. (1983). *Achievement Testing: Recent Advances*. Beverly Hills: Sage Publications.

Bernstein, I. H., and Teng, G. (1989). Factoring items and factor scales are different: Spurious evidence for multidimensionality due to item categorizations. *Psychological Bulletin, 105*, 467-477.

Bock, R. D., Gibbons, R., and Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement, 18*, 155-170.

Drasgow, F., and Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item response. *Journal of Applied Psychology, 68*, 363-373.

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

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gibbons, R. D., Clark, D. C., Cavanaugh, S. V., and Davis, J. M. (1985). Applications of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research, 19*, 43-55.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53*, 525-546.

Glas, C. A. W., and Verhelst, N. D. (1995a). Testing the Rasch model. In Fischer, G. H., and Molenaar, I. W. (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69-95). New York: Springer-Verlag.

Glas, C. A. W., and Verhelst, N. D. (1995b). Tests of fit for polytomous Rasch models. In Fischer, G. H., and Molenaar, I. W. (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 325-352). New York: Springer-Verlag.

Hambleton, R. K., and Swaminathan, H. (1985). *Item Response theory: Principles and Applications.* Boston: Kluwer-Nijhoff Publishing.

Harrison, D. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Multivariate Behavioral Research, 19*, 49-78.

Hsu, T. C., and Yu, L. (1989). Using computers to analyze item response data. *Educational Measurement: Issues and Practice, 8*, 21-28.

Hulin, C. L., Drasgow, F., and Parsons, L. K. (1983). *Item response theory.* Homewood: Dow Jones-Irwin.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*, 223-245.

Linacre, J. M. (1992). Prioritizing misfit indicators. *Rasch Measurement Transactions, 9*, 422-423.

Linacre, J. M. (1998a). Structure in Rasch residuals: Why principal component analysis? *Rasch Measurement Transactions, 12*, 636.

Linacre, J. M. (1998b). Detecting multidimensionality: Which residual data-types works best? *Journal of Outcome Measurement, 2*, 266-283.

Linacre, J. M. (2001). *WINSTEPS* [Computer program, version 3.21]. Chicago: MESA Press.

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

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Schumaker, R. E., Smith, R. M., and Bush, M. J. (2000). Examining replication effects in Rasch fit statistics. In M. Wilson and G. Engelhard, Jr. (Eds.), *Objective Measurement: Theory into Practice* (Volume 5, pp. 303-317). Stamford, CT: Ablex Publishing Corporation.

Slinde, J. A., and Linn, R. L. (1979). The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement, 3*, 437-452.

Smith, Jr., E. V. (2001). Evidence for the reliability of measures and the validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281-311.

Smith, Jr., E. V., and Johnson, B. D. (1998). Factor structure of the DSM –IV criteria for college students using the Adult Behavior Checklist. *Measurement and Evaluation in Counseling and Development, 31*, 164-183.

Smith, Jr., E. V., and Johnson, B. D. (2000). Attention Deficit Hyperactivity Disorder: Scaling and standard setting using Rasch measurement. *Journal of Applied Measurement, 1*, 3-24.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*, 541-565.

Smith, R. M. (1992, April). *Assessing unidimensionality for the Rasch rating scale model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3*, 25-40.

Smith, R. M. (2000). Fit analysis in latent trait models. *Journal of Applied Measurement, 1*, 199-218.

Smith, R. M., and Miao, C. Y. (1991, April). *Assessing unidimensionality for Rasch measurement.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Smith, R. M., and Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice* (Volume 2, pp. 316-327). Norwood, NJ: Ablex.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26*, 249-269.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123-139.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, *3*, 3-24.

Wright, B. D., Mead, R. J., and Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model: Research memorandum No. 22*. Chicago: MESA Press.

Wright, B. D., and Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23-48.

Wright, B. D., and Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

Appendix

Item content of the Adult Behavior Checklist - Revised

1. You fail to pay close attention to details in school.
2. You fail to pay close attention to details at work.
3. You make careless mistakes in school.
4. You make careless mistakes at work.
5. You have difficulty sustaining your attention to tasks/activities.
6. You do not listen when directly spoken to.
7. You do not follow through on instructions.
8. You fail to finish schoolwork.
9. You fail to finish chores.
10. You fail to finish work duties.
11. You have difficulty organizing tasks/activities.
12. You avoid tasks that require sustained mental effort (e.g. homework, schoolwork).
13. You dislike tasks that require sustained mental effort (e.g. homework, schoolwork).
14. You are reluctant to engage in tasks that require sustained mental effort (e.g. homework, schoolwork).
15. You lose things necessary for tasks/activities (e.g. books, keys, tools, school assignments).
16. You are easily distracted by extraneous stimuli (e.g. traffic noises, conversations, looking out the window).
17. You are forgetful in daily activities.
18. You have difficulty remaining quiet during leisure activities.
19. You act as if you are "on the go".
20. You act as if "driven by a motor".
21. You talk excessively.
22. You blurt out answers before questions have been completed.
23. You have difficulty awaiting your turn.
24. You interrupt on others (e.g. butt into conversations or activities).