CA AND SPOD FOR THE ANALYSIS OF TESTS COMPRISED OF BINARY ITEMS

MICHAEL D. MARAUN, JEREMY S. H. JACKSON,
CRAIG R. LUCCOCK, AND SHARON E. BELFER
Simon Fraser University

ROLAND D. CHRISJOHN
Treaty 7 Tribal Council

A test comprised of binary items has an associated theoretical structure (TS) $T_p(D, R, E)$, which may be paired with any of a number of possible quantitative characterizations (QCs). The aim of test analysis is to examine whether a set of test items conforms to an appropriately chosen QC. In this article, the authors consider one of the most common TSs, $T_p(D, R, E)$, in which $D = 1$ construct, R = monotone increasing item/construct regressions, and E = errors in variables. The authors then consider two QCs appropriate for this TS and review tests for each. Several examples are provided.

It is well known that the regression of a dichotomous item on a continuous latent variable is necessarily nonlinear (McDonald, 1980; Mislevy, 1986). Hence, the fitting of a linear model to a set of dichotomous items results in misspecification that undermines the making of meaningful claims about the measurement characteristics of the items (e.g., dimensionality) (Mislevy, 1986). That is, the possibility of initial misspecification means that it is not possible to distinguish between a lack of fit resulting from poor scale performance and a lack of fit resulting from the incorrect choice of a *class* of models to investigate this performance. Conversely, the *adequate* fit of a model to the data in t dimensions might, nevertheless, be an inappropriate

This research was supported in part by a President's Research Grant awarded to the first author by Simon Fraser University. Correspondence should be addressed to Michael Maraun, Department of Psychology, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6; fax: (604) 291-3427.

Educational and Psychological Measurement, Vol. 58 No. 6, December 1998 916-928

representation if misspecification error is large. A more appropriate class of model might require only s ($s < t$) dimensions. This is precisely what occurs when difficulty factors (cf. McDonald & Ahlawat, 1974) arise in the linear factor analysis of dichotomous Guttman scalable items (see, e.g., McDonald & Ahlawat, 1974). In fact, as McDonald (1983) showed, an s-dimensional general nonlinear factor analysis model always is equivalent to a $p \geq s$ dimensional linear factor analysis model. That is, if $\underline{\mathbf{Y}}$ is a p vector of continuous random variables and

$$\underline{\mathbf{Y}} = \underline{\mathrm{m}}(\theta) + \underline{\mathbf{e}},\ E(\underline{\mathbf{Y}}|\theta) = \underline{\mathrm{m}}(\theta),\ E\theta = 0,\ \sigma^2_\theta = 1,\ C(\underline{\mathbf{Y}}|\theta) = \Omega_{\mathrm{diag}},$$

with $\underline{\mathrm{m}}$ representing a p vector of possibly nonlinear functions (and possibly different across items), then

$$C(\underline{\mathbf{Y}}) = C[\underline{\mathrm{m}}(\theta), \underline{\mathrm{m}}(\theta)'] + \Omega_{\mathrm{diag}} = AA' + \Omega_{\mathrm{diag}},$$

with A being a $p \times m$ matrix of real coefficients of rank $= m \geq 1$. Therefore, at the covariance structure level, it is not possible to distinguish between, e.g., a 1-dimensional monotone factor analysis solution and an $s > 1$ dimensional linear factor analysis solution.

However, what seems to be less well respected is the principle that for a scale to be judged as "poor," it must lack empirical conformity to its theoretical structure (TS). Let T_p represent a scale comprised of p dichotomous items and assume that the test was constructed to conform to a particular TS. The TS of a test is, generally speaking, a loose, linguistic specification of how the test is structured. Among other things, it represents, in prematematical terms, how the items are linked to the construct they are designed to measure.

In practice, the TS often lacks formalization and is conceptualized in figurative terms (e.g., these p items *tap* a single *underlying* dimension). On the other hand, it might be implied by the scoring rule of the test (e.g., a total score is computed for T_p, implying a unidimensional TS). Regardless, in most cases, the TS may be represented (somewhat incompletely) as $T_p(D, R, E)$, in which D is the number of constructs that the test items are designed to measure (i.e., the theoretical dimensionality of the test), R is the (theoretical) brand of item/construct regressions, and E is the (theoretical) error structure of the items (e.g., whether they are viewed as "pure" or fallible indicators of the construct to be measured). The TS of a particular test is given by specifying a value for each of D, R, and E.

The TS of a test, while representing in rough terms how the test should behave, has no empirical implications for the multivariate distribution of test items. Hence, the empirical analysis of the performance of a test requires that its $T_p(D, R, E)$ be given a quantitative characterization (QC). A QC is a testable psychometric phrasing of the TS. It is a set of testable requirements for

the multivariate distribution of the test items that, in addition, is "in keeping" with $T_p(D, R, E)$. For any $T_p(D, R, E)$, there will exist many possible QCs. Standard test theory models are QCs of particular $T_p(D, R, E)$. Common unidimensional linear factor analysis, for example, is a possible QC for $T_p(1,$ linear, errors in variables) and, of course, implies that for some latent variable $\boldsymbol{\theta}$, $C(\underline{\mathbf{X}}|\boldsymbol{\theta}) = \Psi_{diag}$, or, equivalently, that the "tetrad condition" holds: $\rho_{xy}{}^*\rho_{uv} = \rho_{xv}{}^*\rho_{uy}$ ($\forall$ x, y, u, v distinct) (Spearman, 1927).

In a "decision-oriented" test analysis, the aim is to arrive at justifiable conclusions about a test's performance and, in particular, to make a decision about whether it is performing "adequately." No claim is made here about whether this aim is a reasonable one, only that this brand of test analysis is common in the social sciences. One examines whether the consequences embodied in a QC hold for the empirical distribution of the test items and speaks of the test as "performing poorly" to the extent that they do not. As an aside, we echo Thissen, Steinberg, Pyszczynski, and Greenberg (1983) in viewing considerations of test score reliability as subordinate to a demonstration that the test is empirically in keeping with its TS.

Clearly, the mere fitting of even an item response model to a set of dichotomous items does not necessarily provide a basis for drawing conclusions about whether the items perform well as a test. The making of sound claims about a test requires a serious consideration of the conformity of the empirical distribution of the items to a properly chosen QC. Now, the TS of many tests comprised of dichotomous tests implies nothing more than monotone increasing regressions of the items on a single implied latent variable, that is, something like $T_p(1,$ monotone increasing, errors in variables).[1] The regressions are not required to be of any particular parametric form, just monotone. For such tests, decision-oriented test analysis is not possible with generic item response models. In this article, we review useful but little-used theory for tests with TSs of this type and provide several examples illustrating these procedures.

Unidimensional Monotone Latent Variable and Monotone Positively Correlated Latent Variable Models: Two Quantitative Characterizations

An appropriate choice of QC for tests with $T_p(1,$ monotone increasing, errors in variables) is the class of unidimensional monotone latent variable (UMLV) models (Holland & Rosenbaum, 1986). UMLV models are item response models,

$$P(\underline{\mathbf{X}} = X_\bullet) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_j P(X_j = 1|\underline{\boldsymbol{\theta}})^{X_{j\bullet}} [1 - P(X_j = 1|\underline{\boldsymbol{\theta}})]^{1-X_{j\bullet}} dF(\underline{\boldsymbol{\theta}}), \quad (1)$$

with $\underline{\mathbf{X}}$ a p vector of 0/1 random variates and with the following specializations: θ is a scalar, the item/latent variable regression $P(\mathbf{X}_j = 1|\theta)$ is any increasing function of θ (possibly different across items), and $F(\theta)$ is arbitrary. UMLV models are unidimensional in the standard sense of conditional independence given the latent variable θ. A second class of characterizations that seems particularly useful is the class of monotone positively correlated latent variable (MPCLV) models. These models are a specialization of Model 1, in which there are $m > 1$ positively correlated latent variables, and the regression of each item on each latent variable is monotone increasing. Put another way, MPCLV models do not include the specification of conditional independence given a single θ. For obvious reasons, this class of models is a less pleasing reality for a set of dichotomous items. As Holland and Rosenbaum (1986) stated, a scalar θ "easily lends itself to the interpretation as an underlying ~'true' quantity that is fallibly measured by the observable responses in $\underline{\mathbf{X}}$" (p. 1526). Unidimensionality also provides a justification for summative scoring rules. However, even a well-constructed test with a unidimensional TS may, in practice, include a number of positively related subclusters of items. Thus, in certain contexts, an MPCLV characterization can be viewed as a somewhat unsuccessful attempt to produce a unidimensional test.

Holland and Rosenbaum (1986) showed that if j dichotomous items are described by a UMLV model, then they are conditionally associated (CA), and certain relations hold among the set of 2^j observed proportions, $P(\underline{\mathbf{X}} = \underline{X}_i)$. In particular, the conditional covariances of all nondecreasing functions of any subset of the j items, given any function of the remaining items, are non-negative. That is,

$$C[d(\underline{\mathbf{Y}}), g(\underline{\mathbf{Y}})|h(\underline{\mathbf{Z}})] \geq 0 \ \forall \ d, g \text{ nondecreasing}, \underline{\mathbf{X}}' = (\underline{\mathbf{Z}}', \underline{\mathbf{Y}}').$$

The non-negativity of covariances of nondecreasing functions of the items clearly is a special case of CA, in which the conditioning is on the empty set. If it is found that the items are, within reason, CA, then (if desired) a search may be undertaken for a particular UMLV representation for the items. If, on the other hand, it is found that the items are not CA, then it may be concluded that no UMLV model describes the j items and that the items are not unidimensional in this broad sense. Interestingly, it is not clear whether CA is sufficient for a UMLV representation (Holland & Rosenbaum, 1986).

On the other hand, a set of items that is described by a model within the class of MPCLV models must exhibit strong positive orthant dependence (SPOD) (Holland, 1981; Joag-Dev, 1983). For dichotomous items, this means that for all pairs of disjoint subsets, A (with k items) and B (with n items), of the j items (with $k + n \leq j$),

$$P(\underline{\mathbf{X}}_A = \underline{1} \cap \underline{\mathbf{X}}_B = \underline{1}) \geq P(\underline{\mathbf{X}}_A = \underline{1}) \, P(\underline{\mathbf{X}}_B = \underline{1}), \tag{2}$$

$$P(\underline{X}_A = \underline{0} \cap \underline{X}_B = \underline{0}) \geq P(\underline{X}_A = \underline{0})\, P(\underline{X}_B = \underline{0}), \text{ and} \tag{3}$$

$$P(\underline{X}_A = \underline{1} \cap \underline{X}_B = \underline{0}) \leq P(\underline{X}_A = \underline{1})\, P(\underline{X}_B = \underline{0}). \tag{4}$$

where $\underline{1}$ is a vector of 1^s and $\underline{0}$ is a vector of 0^s. If the items are SPOD, then (if desired) a search may be undertaken for an appropriate representation from the class of MPCLV models. Holland (1981) showed that SPOD is a sufficient condition for the existence of an MPCLV representation for the items.

Although SPOD and CA are complete test conditions, in application they are not without their problems (Zwick, 1986). In particular, the sets of conditions that comprise SPOD and CA are too large to test comprehensively. Therefore, the investigator must choose a subset of conditions and settle for an incomplete test. This introduces a problem with the power of each test. If the items do not conform to the chosen subset of conditions, then one may reject CA (SPOD). On the other hand, if the items do conform, then one may merely lack the power to correctly reject CA (SPOD). The power issue is then 2-dimensional. It pertains to both the proportion of conditions one requires to make a reasonable decision about whether the items are CA or SPOD, and statistical sample size requirements. Exactly how large a proportion of conditions (and which conditions) is required is an interesting question, one not dealt with in the current work. Regardless, it might be argued that it is more desirable to assess whether a set of items is *roughly* CA (SPOD) via an incomplete test than to make a rash conclusion on the basis of an inappropriate QC.

It also might be asked why one would not simply employ any of a number of common procedures, for example, Stout's (1987) DIMTEST theory, the linear factor analysis of a matrix of tetrachoric correlations, or parametric item response theory. The reason is that there is more than just one definition of unidimensionality, and these approaches do not instantiate the particular brand of unidimensionality implied by T_p(1-dimensional, monotone increasing, errors in variables). Let $\underline{\mathbf{Y}}$ be a p vector of continuous latent variates and $\underline{\mathbf{X}}$ be obtained by dichotomizing $\underline{\mathbf{Y}}$ according to a p vector of thresholds $\underline{\gamma}$. Then both the linear factor analysis of a matrix of tetrachoric correlations and Model 1 with $P(\underline{\mathbf{X}}_j = 1|\theta) = \Phi[a_j(\theta - b_j)]$ attempt to assess whether $\underline{\mathbf{Y}}$ is unidimensional in a *linear* factor analytic sense. That is, they attempt to determine whether there is a random variable $\boldsymbol{\theta}$ (the common factor), with $E(\boldsymbol{\theta}) = 0$ and $\sigma^2_{\theta} = 1$ such that

$$(\underline{\mathbf{Y}}|\theta) \sim N(\underline{\Lambda}\theta, \Psi_{\text{diag}}), \text{ and } \psi_j = (1 - \lambda^2_j),$$

so that $R_{\underline{Y}} = \underline{\Lambda\Lambda}' + \Psi$.

This is clear from the following argument. Consider the parametric case of Model 1 with $P[\mathbf{X}_j = 1|\theta) = \Phi(a_j(\theta - b_j)]$, that is, the normal ogive random item response model,

$$P(\underline{X} = \underline{X}_*) = \int_{-\infty}^{\infty} \prod_j [\Phi(a_j(\theta - b_j))]^{X_j} [1 - \Phi(a_j(\theta - b_j))]^{1-X_j} dF(\theta). \quad (5)$$

Model 5 is a representation of $\underline{\mathbf{X}}$ if and only if $\underline{\mathbf{X}}$ has an equivalent representation

$$P(\underline{X} = \underline{X}_*) = \int_{-\infty}^{\infty} \int_{\Psi}^{\infty} f(\underline{Y}|\theta)\, d\underline{Y} dF(\theta) = \int_{\Psi}^{\infty} f(\underline{Y})\, d\underline{Y} \quad (6)$$

with $\underline{\mathbf{Y}}$ a p vector of continuous latent variates, $\underline{\gamma}$ a p vector of thresholds, and

$$(\underline{\mathbf{Y}}|\theta) \sim N(\underline{\Lambda}\theta, \Psi_{diag}),\ \sigma^2_\theta = 1, \text{ and } \psi_j = (1 - \lambda^2_j).$$

That is, $\underline{\mathbf{X}}$ is representable as in Model 5 if and only if it is representable as the dichotomization of a vector of latent variates $\underline{\mathbf{Y}}$ that conform to a common unidimensional linear factor model (Bartholomew, 1981; Maraun, 1993; Muthen, 1978). In particular, $f(\underline{\mathbf{Y}}|\theta)$ must be multivariate normal, the $\mathbf{Y}_j$ must be independent conditional on θ, and $E(\underline{\mathbf{Y}}|\theta)$ must be a linear function of θ. In this case, $\Sigma_{\underline{Y}} = R_{\underline{Y}} = \underline{\Lambda\Lambda'} + \Psi$ and contains the correlations estimated by the tetrachoric correlations computed on the $\mathbf{X}_j$. The following parameter correspondence exists between the two representations:

$$\lambda_j = \frac{a_j}{\sqrt{1 + a_j^2}}, \quad \gamma_j = \frac{a_j b_j}{\sqrt{1 + a_j^2}}.$$

On the other hand, CA tests whether there is a random variable $\boldsymbol{\theta}$ (the common factor), with $E(\boldsymbol{\theta}) = 0$ and $\sigma^2_\theta = 1$ such that

$$f(\underline{\mathbf{Y}}|\theta = \theta_o) = \Pi f(\mathbf{Y}_j|\theta = \theta_{oj}) \quad (7)$$

$$E(\underline{\mathbf{Y}}|\theta) = \underline{m}(\theta) \text{ not necessarily linear, but monotone increasing.} \quad (8)$$

Hence, the aim of CA also is to assess whether $\underline{\mathbf{Y}}$ is unidimensional but, quite clearly, in a different sense than the unidimensionality embodied by linear factor analysis. The failure to instantiate the particular brand of unidimensionality called for by T_p(1, monotone increasing, errors in variables) also is why DIMTEST is not an appropriate choice.

An (incomplete) test of CA may be carried out in two stages. In Stage 1, the item covariances are checked for non-negativity. Given that CA is not rejected in Stage 1, a strong (but incomplete) test of CA can be made, following Zwick (1986) and Holland and Rosenbaum (1986), with Mantel-Haenszel test statistics (Mantel & Haenszel, 1959). Specifically, the conditional covariance of each pair of items, given the total score on the remaining $(p - 2)$ items, is tested for non-negativity. The Mantel-Haenszel statistic, in this case, is a weighted average of the conditional covariances in each 2×2 slice of the

$2 \times 2 \times (p-2)$ contingency table of item pair by total score. Due to the large number of tests typically involved, α should be set to a conservative value. To be on the safe side, 5% or more of the individual tests should indicate a statistically significant negative conditional covariance before declaring that the test as a whole does not exhibit CA. The test of SPOD is likewise limited by the heavy computational cost of dealing with the full set of inequalities pertinent to SPOD. Therefore, the search for disconfirming evidence may begin with 3-way subsets of the full p-way array. For 3-way data, inequalities of Type 3, for example, are of the form

$$P(\text{ITEM}_1 = 1 \cap \text{ITEM}_2 = 1 \cap \text{ITEM}_3 = 1) \geq P(\text{ITEM}_1 = 1 \cap \text{ITEM}_2 = 1)\, P(\text{ITEM}_3 = 1).$$

The rejection rule for SPOD can be set as for CA, that is, if more than 5% of the inequalities do not hold.

Example 1: Data From the Self-Monitoring Scale

To date, many analyses of the Self-Monitoring Scale (SMS) (Snyder, 1974) have been carried out. The conclusions arising from these studies have not been very positive, for although the scale has a unidimensional TS, the majority of these studies (Briggs & Cheek, 1988; Briggs, Cheek & Buss, 1980; Hoyle & Lennox, 1991; Tobey & Tunnell, 1981) have shown the scale to be of a higher dimensionality. The consensus now seems to be that the scale has a dimensionality of at least three. If a disparity between the TS and empirical dimensionality of the SMS does in fact exist, then it has implications for the scale's use. For example, such a disparity suggests that the summative scoring rule of the SMS lacks justification. This very possibility has, in fact, motivated a number of major revisions to the scale (e.g., Gangestad & Snyder, 1985; Lennox & Wolfe, 1982).

The conclusion that the scale does not perform as it should might, however, be somewhat rash. This is because the majority of previous studies have employed linear factor and component models to analyze the SMS. Yet, these models do not provide a basis for reaching justifiable conclusions about the scale's possible departure from its TS. In the first place, the SMS is comprised of dichotomous items. In the second place, the TS of the SMS implies nothing more than monotone increasing regressions. For these two reasons, it is not appropriate to pass judgment on the performance of the SMS based on linear factor models or even generic item response models. An appropriate QC for the SMS is the class of UMLV models. Empirical conformity to a UMLV characterization would justify the summative scoring rule of the SMS. A second, less adequate candidate is MPCLV.

Responses to the 25 items of the SMS were obtained from a sample of 903 students at the University of Guelph. Items and their means are provided in

Table 1. An examination of the 300 covariances revealed that there were only a small number of (marginally) negative covariances. For the Mantel-Haenszel statistics, α was set to .01 with a resulting one-tailed critical value of 2.33. Of the 300 Mantel-Haenszel statistics computed, 116 (38.7%) were negative, 10% leading to rejection of the non-negativity hypothesis. Therefore, it was concluded that the scale did not have a UMLV representation and so was not unidimensional in this sense. For the test of SPOD, 45.6% of the required 3-way inequalities were violated. Therefore, it was concluded that the scale was not in keeping with an MPCLV representation. What is implied by these findings is that at least some of the items have nonmonotone relations with the latent variables or, in other words, are characterized by multiple negatively correlated latent variables. This structuring of the items obviously is a serious departure from the TS of the SMS.

Example 2: Artificial Data

This section is not in any way presented as a simulation study. Instead, the section presents the analyses of several sets of data useful in illustrating points made earlier in the article. Consider Model 6 with $(\underline{\mathbf{Y}}|\theta) \sim N(\underline{\mathrm{m}}(\theta), \Psi_{\mathrm{diag}})$, $\theta \sim N(0, 1)$; and $m_1 = \exp[.24(\theta - .75)] / 1 + \exp[.24(\theta - .75)]$; $m_2 = .5\ln(\theta + 4)$; $m_3 = \exp(\theta^3)$; $m_4 = \exp[5.05(\theta - 1.75)] / 1 + \exp[5.05(\theta - 1.75)]$; $m_5 = \exp[.75(\theta - 1.75)] / 1 + \exp[.75(\theta - 1.75)]$; $m_6 = .001\theta^3 + .05\theta$; $m_7 = .005\ln(\theta + 3)$; $\gamma' = [.5\ .35\ .15\ 1.4\ -.3\ -.25\ .2]$; $\psi_j = .04\ \forall\, j$. The regressions are depicted in Figure 1. With these choices for $\underline{\mathrm{m}}(\theta)$, this is a UMLV model. According to the previously reviewed theory, a program such as TESTFACT (Wilson, Wood, & Gibbons, 1991), which employs full information factor analysis to fit Model 5, should not lead to a correct decision about data generated on the basis of this UMLV model. A test of CA, on the other hand, should lead to the correct conclusion that data generated on the basis of this model are describable within the class of UMLV models. A total of 900 realizations were taken of $\underline{\mathbf{X}}$, generated on the basis of this UMLV model. TESTFACT rejected solutions of from one to three dimensions, whereas all 21 Mantel-Haenszel statistics were positive, leading to the correct decision that $\underline{\mathbf{X}}$ was representable within the class of UMLV models.

Consider a 2-dimensional version of Model 6 for 25 items in which $(\underline{\mathbf{Y}}|\underline{\theta}) \sim N(\Lambda\theta, \Psi_{\mathrm{diag}})$, $\underline{\theta} \sim N_2(\underline{0}, \mathrm{I})$, $\psi_j = (1 - \lambda^2_j)$, and γ_j range from -1.5 to $+1.5$. In addition, for 13 of the items $\lambda_{j1} = .8$ and $\lambda_{j2} = .1$, whereas for the other 12 items $\lambda_{j1} = .1$ and $\lambda_{j2} = .8$. Hence, $\Sigma_{\underline{Y}} = R_{\underline{Y}} = \Lambda\Lambda' + \Psi$. This model is not a UMLV model but is instead an MPCLV model. A total of 900 realizations of $\underline{\mathbf{X}}$ were generated on the basis of this model. The Mantel-Haenszel procedure was then employed to test whether the items were CA. Figure 2a is a histogram for the 300 Mantel-Haenszel statistics.

Table 1
Means for Items of Self-Monitoring Scale

Item	Mean
1. I find it hard to imitate the behavior of other people (F)	.421
2. My behavior usually is an expression of my true inner feelings, attitudes, and beliefs (F)	.344
3. At parties and social gatherings, I do not attempt to do or say things that others will like (F)	.301
4. I can only argue for ideas that I already believe (F)	.476
5. I can make impromptu speeches on topics about which I have almost no information (T)	.316
6. I guess I put on a show to impress or entertain people (T)	.340
7. When I am uncertain how to act in social situations, I look to the behavior of others for cues (T)	.811
8. I probably would make a good actor (T)	.344
9. I rarely need the advice of my friends to choose movies, books, or music (F)	.737
10. I sometimes appear to be experiencing deeper emotions than I actually am (T)	.332
11. I laugh more when I watch a comedy with others than when I am alone (T)	.464
12. In a group of people, I rarely am the center of attention (F)	.565
13. In different situations and with different people, I often act like very different people (T)	.585
14. I am not particularly good at making other people like me (F)	.163
15. Even if I am not enjoying myself, I often pretend to be having a good time (T)	.403
16. I am not always the person I appear to be (T)	.644
17. I would not change my opinions (or the way in which I do things) to please someone else or win their favor (F)	.615
18. I have considered being an entertainer (T)	.212
19. To get along and be liked, I tend to be what people expect me to be rather than anything else (T)	.184
20. I have never been good at games like charades or improvisational acting (F)	.457
21. I have trouble changing my behavior to suit different people and different situations (F)	.236
22. At parties, I let others keep the jokes and stories going (F)	.520
23. I feel a bit awkward in company and do not show up quite as well as I should (F)	.335
24. I can look anyone in the face and tell a lie with a straight face (if for the right end) (T)	.478
25. I may deceive people by being friendly when I really dislike them (T)	.542

Note. The symbol T (F) indicates that endorsement (lack of endorsement) of the item is scored in the direction of self-monitoring.

With α set (as before) to .01, 154 of the 300 tests resulted in rejections, leading to the decision that the items were not describable as a UMLV model. It also is interesting to observe the complete separation of the distributions of negative and positive Mantel-Haenszel statistics. It is our experience that for tests with a large number of items, a histogram of Mantel-Haenszel statistics is very useful in diagnosing the structure of the items (much as

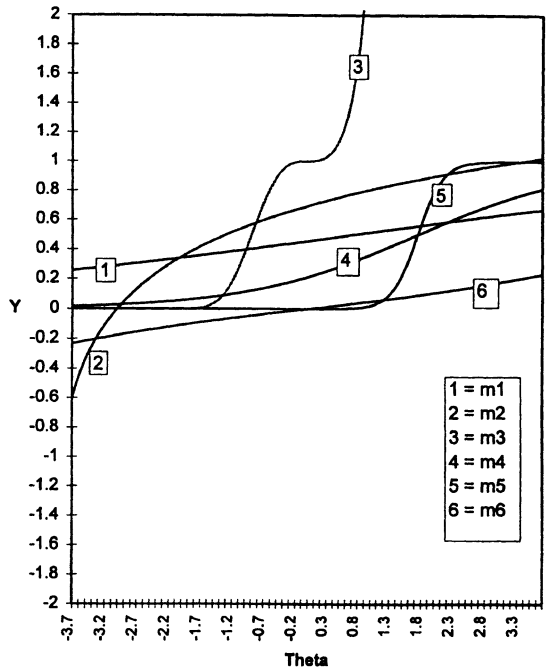

Figure 1. Regressions of $\mathbf{Y}_j$ on $\boldsymbol{\theta}$.

autocorrelation function plots are useful in identifying time-series models). For the test of SPOD, 1.9% of the required inequalities were violated, leading to the conclusion that the data are indeed describable in MPCLV terms.

As a final example, consider Model 6, for 25 items, in which $(\underline{\mathbf{Y}}|\boldsymbol{\theta}) \sim N(\underline{\Lambda}\boldsymbol{\theta}, \boldsymbol{\Psi}_{\text{diag}})$, $\boldsymbol{\theta} \sim N(0, 1)$, $\psi_j = (1 - \lambda^2_j)$, γ_j range from -1.5 to $+1.5$, $\underline{\Lambda}' = [\underline{\Lambda}_1', \underline{\Lambda}_2', 0]$, $\underline{\Lambda}_1'$ $= [-.9\ -.8\ -.7\ -.6\ -.5\ -.4\ -.3\ -.2\ -.2\ -.1\ -.1]$, and $\underline{\Lambda}_2' = -\underline{\Lambda}_1'$. Obviously, this is neither a UMLV nor an MPCLV model. A total of 900 realizations of $\underline{\mathbf{X}}$ once again were generated on the basis of this model. Figure 2b is a histogram for the Mantel-Haenszel statistics. Approximately 26% of the Mantel-Haenszel tests were statistically significant, leading to the correct rejection of UMLV. Once again, a separation between the distribution of the positive and negative statistics clearly is evident. As one would expect, this separation disappears as negative loadings become less extreme.

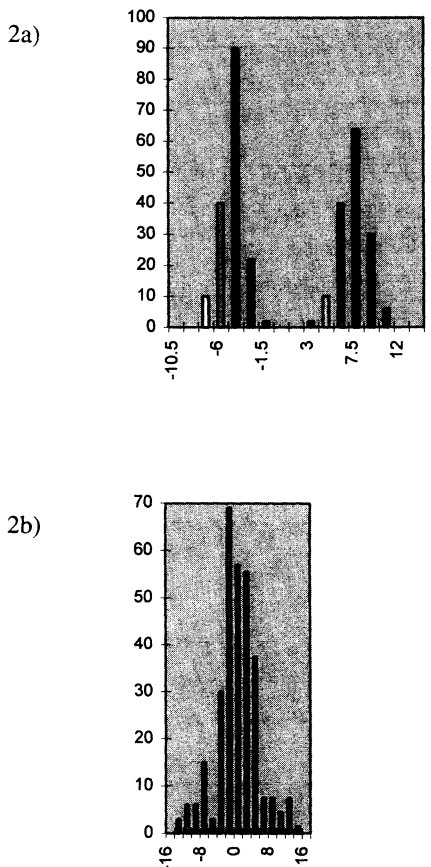

Figure 2. Histograms of Mantel-Haenszel statistics.

Discussion

The aim of this article certainly was not to enshrine the use of Mantel-Haenszel statistics as a new dogma of test theory. Instead, the aim was chiefly to endorse logical test analytic thinking. Specifically, in applied test theory, researchers tend to employ models in a generic fashion. However, a model has implications for the quality of a test only if it is a QC for the TS of the test. Spearman (1927) created linear factor analysis as a mathematical paraphrase

of his ideas about intelligence tests. This is the type of thinking that applied test theory requires, for test theory is fundamentally about the correct paraphrase of TSs. The details associated with the implementation of each particular QC, including that of CA and SPOD, is an important but secondary matter.

For the particular case of tests comprised of dichotomous items and with T_p(1, monotone increasing, errors in variables), the classes of UMLV and MPCLV models are arguably appropriate QCs. Judging from the applied literature, this fact does not seem to be well known. What one sees in practice are "test analyses" in which a parametric item response model or, even worse, a linear factor model is applied to items regardless of the particularities of their TS. It is suggested here that this type of analysis is somewhat antithetical to the purpose of test analysis. The purpose of test analysis is not typically the mere representation of a set of items by a mathematical model but rather the assessment of whether the test conforms to its TS. This clearly was the aim of previous analyses of the SMS. For example, Hoyle and Lennox (1991) state, "Although the construct of self-monitoring has assumed a central role in the description and explanation of human behavior, there is considerable disagreement about the performance of the Self-Monitoring Scale" (p. 511). Given this aim, it is unfortunate the frequency with which test analyses are rendered impotent by a lack of care in the pairing of QC with TS.

Note

1. It is assumed throughout that the items have been recoded in such a way that they all have monotone increasing regressions.

References

Bartholomew, D. (1981). Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology, 34*, 93-99.

Briggs, S., & Cheek, J. (1988). On the nature of self-monitoring: Problems with assessment, problems with validity. *Journal of Personality and Social Psychology, 54*, 663-678.

Briggs, S., Cheek, J., & Buss, A. (1980). An analysis of the Self-Monitoring Scale. *Journal of Personality and Social Psychology, 38*, 679-686.

Gangestad, S., & Snyder, M. (1985). "To carve nature at its joints": On the existence of discrete classes in personality. *Psychological Review, 92*, 317-340.

Holland, P. (1981). When are item response models consistent with observed data? *Psychometrika, 46*, 79-82.

Holland, P., & Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14*, 1523-1543.

Hoyle, R., & Lennox, R. (1991). Latent structure of self-monitoring. *Multivariate Behavioral Research, 26*, 511-540.

Joag-Dev, K. (1983). Independence via uncorrelatedness under certain dependence structures. *Annals of Probability, 11*, 1037-1041.

Lennox, R., & Wolfe, R. (1982). Revision of the Self-Monitoring Scale. *Journal of Personality and Social Psychology, 46*, 1349-1364.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Maraun, M. (1993). *Issues pertaining to the determinacy of item response models*. Unpublished doctoral thesis, University of Toronto.

McDonald, R. (1980). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

McDonald, R. P. (1983). Exploratory and confirmatory nonlinear factor analysis. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A Festschrift for Frederic M. Lord*. Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R., & Ahlawat, K. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99.

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

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551-560.

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30*, 526-537.

Spearman, C. (1927). *Abilities of man*. New York: Macmillan.

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

Thissen, D., Steinberg, L., Pyszczynski, T. & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7*, 211-226.

Tobey, E., & Tunnell, G. (1981). Predicting our impressions on others: Effects of public self-consciousness and acting—A Self-Monitoring subscale. *Personality and Social Psychology Bulletin, 7*, 661-669.

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

Zwick, R. (1986, April). *Assessing the dimensionality of dichotomous item responses: Theoretical and empirical perspectives*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.