A Multisample Item Response Theory Analysis of the Beck Depression Inventory-1A

ROBINDER P. BEDI, MICHAEL D. MARAUN, *Simon Fraser University,*
and ROLAND D. CHRISJOHN, *St. Thomas University*

Abstract

The widespread employment of the Beck Depression Inventory-1A (BDI-1A) has spawned a number of practices: 1) The employment of an unweighted total score as a measure of depression; 2) Its use in populations other than that in which it was normed; and 3) The employment of BDI-1A total scores in hypothesis tests about population differences in mean depression. A sequential procedure based on item response theory was employed to assess the validity of these practices for the case of four populations: clinical depressives, mixed nondepressed psychiatric patients, and students from two different universities. The findings suggested that the first practice was not justified for any of these populations, that the BDI-1A was employable only with clinical depressives and with one of the university populations, and that mean comparisons were not allowable.

Résumé

L'emploi largement répandu de l'inventaire de dépression de Beck – Version 1A (IDB-1A) a engendré de nombreuses pratiques : 1) l'utilisation des scores obtenus comme mesures de la dépression ; 2) le recours à l'IDB auprès de populations différentes de celles pour qui ce test a été conçu ; et 3) l'application des scores totaux obtenus à l'IDB-1A à des tests d'hypothèses portant sur les différences de l'indice moyen de dépression dans une population. On a évalué la validité de telles pratiques au moyen d'une procédure séquentielle basée sur la théorie de la réponse d'item auprès de quatre populations différentes : des personnes souffrant de dépression clinique, des patients en établissement psychiatrique souffrant ou non de dépression, et des étudiants de deux universités distinctes. Les conclusions de l'étude révèlent que le recours à la première pratique n'était justifié auprès d'aucune des populations, que l'usage de l'IDB-1A n'était approprié qu'auprès de la population composée de personnes souffrant de dépression clinique et que les comparaisons des indices moyens de dépression n'étaient pas acceptables.

The Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erlbaum, 1961) is a self-report inventory developed to assess the "depth" of depression of individuals already diagnosed as depressed (Beck, 1972, p.187). Each of the twenty-one items corresponds to a particular, putative symptom of depression, and is paired with a 4-point Likert response scale. In 1979, Beck, Rush, Shaw, and Emery published a modified version of the original Beck in which alternative wordings for the same symptoms, and double negatives, were eliminated. This revised version, the BDI-1A, has become one of the most popular instruments for the assessment of depression (Piotrowski & Keller, 1992).

The widespread employment of the BDI-1A in both research and applied settings has spawned a number of related practices, among these: 1) The employment of a simple, unweighted composite of BDI-1A items (i.e., total score) as a measure of depression; 2) The employment of the BDI-1A in populations other than that in which it was normed; 3) The use of the BDI-1A total score as input into tests of hypotheses about population mean differences in level of depression. Each of these practices is justified by specific kinds of psychometric evidence. For while the scoring of a test as a single composite of the items, and, in particular, an unweighted linear composite, is, in certain quarters, virtually the default, this practice is justified only given that, within the population of interest, the test items behave psychometrically in a particular way. Usually, the requirement is that the items manifest a particular brand of unidimensionality. With respect to the second practice, the BDI-1A was normed in a population comprised of clinically depressed in- and outpatients. Assuming that it had been shown to perform satisfactorily within this norming population, its employment in any other population, say population B, would then be justified by evidence of its satisfactory performance in this population. Failure to provide evidence that an instrument performs satisfactorily in contexts beyond those for which it was designed, prior to its employment in such contexts, is in clear

violation of the ethical standards of the American Psychological Association (American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 1985). Finally, assuming that a test possesses satisfactory psychometric properties in two populations, say populations A and B, comparing the means of A and B on any function of the test items is justified by the demonstration that certain cross-population item parameter invariances hold. In the absence of these invariance relations, it is not, in general, possible to distinguish between the scenario of population mean differences in the measured construct, and that of different constructs being measured by the instrument in the two populations (Thissen, Steinberg, & Gerrard, 1986).

As long as psychological tests are to be employed in research and applied settings, it is essential that formal justification be provided for the uses to which they are put. In fact, construct validation theory (e.g., Cronbach & Meehl, 1955) calls for an ongoing program of investigation into the justification of the claims that are made on the basis of a given test. Modern test theory provides a coherent, formal, sequential logic which may be employed to assess whether, within a given context, practices (1), (2), and (3) are justified. In general terms, this strategy may be described as follows:

1) Associated with each test is a theoretical structure (TS). A TS is a loose linguistic specification of how the items of a test are designed to measure, and, in particular, of how they are viewed as linked to the construct that they were designed to measure, and of whether or not they are viewed as "fallible" indicators of the construct. The TS of a test is, on the linguistic plane, the standard of correctness to which the test is compared in a test analysis. Hence, the first step in a test analysis is to specify the theoretical structure of the analyzed test. The TS of a given test may be formalized as a 4-tuple,

$$\text{TS}(I,D,R,E),$$

in which D stands for the number of constructs the items are designed to measure, R, the form of the (theoretical) regressions of the items on the construct that they are designed to measure, and E, the error structure of the items (see Maraun, Jackson, Luccock, Belfer, & Chrisjohn, 1998). The initial variable I merely refers to the type of items (e.g., continuous, dichotomous, 7-point Likert) of which the test is comprised.

2) Being as the TS is a purely linguistic construction, it does not imply any requirements of the empirical behaviour of the test items, in order that the test's performance may justifiably be judged as satisfactory. The generation of empirical test analytic requirements comes about by way of a mathematical paraphrase of the TS. Specifically, for each TS there may be constructed a set of quantitative characterizations (QCs), each QC a mathematical translation of the TS. A QC is a set of quantitative, empirical requirements for the distribution of the test items that is also "in keeping" with the TS. It is, in other words, the quantitative embodiment of the TS, and specifies how the test should behave empirically, if its performance is to be judged as satisfactory. The majority of QCs in existence are, of course, unidimensional models such as the unidimensional, linear factor model, and the unidimensional item response models. Each of these models may be seen as a mathematical paraphrase of a particular TS.

3) A test analysis is, minimally, an assessment of whether a given test conforms to its theoretical structure. Once an appropriate QC is chosen, this assessment is equivalent to an empirical assessment of whether the multivariate distribution of the test items meets the requirements specified by the QC.

4) A scoring rule for a test is a single valued function of the test items, say $\mathbf{t} = f_{sc}(\mathbf{X}_1,\mathbf{X}_2,....,\mathbf{X}_P)$. A test may be justifiably scored only under certain conditions. Generally speaking, a test may be scored if it conforms empirically to an appropriately chosen QC (standardly a unidimensional model of some sort). The QC, in concert with a chosen statistical rationale, implies a specific form for the scoring rule and, hence, determines whether the use of any particular scoring rule, including the unweighted sum, is justified (see, e.g., Thissen, Steinberg, Pyszczynski, & Greenberg, 1983).

5) If, by (4), there exists a scoring rule for a given test, then the reliability of the function thus defined, [i.e., $\mathbf{t} = f_{sc}(\mathbf{X}_1,\mathbf{X}_2,....,\mathbf{X}_p)$], may be examined, and is called the reliability of the test. The specific computational formula is determined by the form of the scoring rule, which, in turn, depends upon the pair of QC and chosen statistical rationale.

6) If, within a given population of respondents, say population A, a test conforms to an appropriately chosen QC and, once scored, is shown to possess adequate reliability, then its performance may, provisionally, be judged as satisfactory within this population. For the test to be employed in any other population, say population B, its performance must be shown to be satisfactory within this population.

7) If a test is shown to perform satisfactorily within both populations A and B, then the means of these populations on variate $\mathbf{t} = f_{sc}(\mathbf{X}_1,\mathbf{X}_2,....,\mathbf{X}_p)$ may justifiably be compared if the psychometric properties of the test are identical in the two populations. In practice,

this will require that the numerical values of the parameters of the chosen QC are invariant over the populations.

Although Beck (1972) did not explicitly delineate the theoretical structure of the BDI-1A, it can, nevertheless, be deduced in part from his writing and, in part, from logical considerations. First, the response scales of the items are 4-point Likert and, hence, should not properly be viewed as continuous. Second, since each item on the BDI-1A is designed to measure the frequency or intensity of a particular depressive symptom (Beck, 1972) and since, "by grouping patients together on the basis of their having scored high...we would have a population that would be congruent so far as this particular variable was concerned..." (Beck, 1972, p. 202), it may be concluded that the BDI-1A was designed to measure but one construct, that being depression. Third, Beck (1972) states that "with increasing severity of depression... there is a step-like progression in the frequency of depressive symptoms, the more depressed a patient is, the more intense a particular symptom is likely to be" (p. 202). In other words, the more severe the depression, the more strongly manifest will be each symptom. This suggests that the item-construct regressions are to be considered monotone increasing (MI). Fourth, as no mention is made to the contrary, it may reasonably be assumed that the items are free to vary in regard to their sensitivity to changes in the severity of depression. Hence, there exists latitude for the parametric form of the monotone increasing item-depression regressions to vary over items. Finally, as with most tests, it is safe to assume that the items are intended to be error-in-variable (EIV) indicators of depression. Hence, equally depressed individuals are not expected to respond in an identical manner to the test items. Jointly, these considerations suggest that the theoretical structure of the BDI-1A is TS(4-point Likert,1,MI,EIV).

In examining the psychometric properties of the BDI-1A, many researchers have employed linear factor analysis (e.g., Golin & Hartz, 1979; Hill, Kemp-Wheeler, & Jones, 1986). It is clear, however, that linear factor analysis, which assumes the items to be continuous variates, and asserts a linear item/construct regression, is an inappropriate QC for a test that is characterized as TS(4-Likert,1,MI,EIV). Hence, the results it generates can have no legitimate bearing on judgments regarding the performance of the BDI-1A. A better quantitative translation of TS(4-Likert,1,MI,EIV) is Samejima's (1969) graded item response model. The BDI-1A is comprised of 21 items, each with a 4-option Likert response scale. As with all item response models, Samejima's graded model describes the probability that a randomly sampled individual will endorse option k of item j, $P(\mathbf{X}_j = k \mid \theta)$, as a function of his position along a latent continuum θ (viewed, in this case, as depression), and a set of item parameters. Specifically, the option characteristic curves are given as

$$P(\mathbf{X}_j = k \mid \theta) = 1/(1 + \exp(-a_j(\theta-\tau_{j(k)})) - 1/(1 + \exp(-a_j(\theta-\tau_{j(K+1)})),$$

in which a_j denotes the slope, and $\tau_{j(k)}$ the kth threshold, parameter of item j (Samejima, 1969). The parameter $\tau_{j(k)}$ is the point on the θ continuum at which the probability passes .50 that a randomly sampled individual will endorse option k or higher. By specifying a distribution for the latent variate θ, the model describes the observed proportions, $P(\boldsymbol{X} = X_*)$, each proportion corresponding to one of the 4^{21} response patterns that may arise when the BDI-1A is administered, as

$$P(\boldsymbol{X} = X_*) = \int_{-\infty}^{\infty} \prod_{j=1}^{J} P(X_j = X_{j*} \mid \theta)\, f(\theta)\, d(\theta)$$

The probability $P(\boldsymbol{X}_j = X_{j*} \mid \theta)$ is the probability, conditional on θ, that item j assumes score X_{j*} (one of the K response options), and $f(\theta)$ is the density function of the latent variate θ. For the BDI-1A, there will be 21 slope parameters and 63 threshold parameters.

If, within a given population, the graded item response model describes the distribution of the items of the BDI-1A, then the test may be scored. The form of the scoring rule will depend upon both the item parameters of the model (i.e., the QC), and the choice of a statistical rationale to estimate the latent variate scores of the individuals under study. Traditionally, for item response models, one chooses the scoring rule that is "good" in the statistical sense of relative efficiency (Lord & Novick, 1968). The maximum likelihood estimator of θ is close to being optimal in this sense. However, it is a nonlinear scoring rule. If the researcher would prefer to score the test as a simple weighted, linear composite of the items, there are two possibilities. If the slopes, a_j, are equal over the J items, then the appropriate scoring function is the unweighted item composite or simple total score. The finding of equal slope parameters is, in other words, the condition under which the "default" scoring rule is justified. If, on the other hand, the slopes are not equal over items, the appropriate scoring function is the weighted sum, $\Sigma a_j X_j$. Once a scoring function has been derived for the test, a lower bound to the reliability of the test may then be defined as

$$1 - \frac{E(\sigma^2_{fsc(X1,X2,...,XP)} \mid \theta)}{\sigma^2_{fsc(X1,X2,...,XP)}}$$

in which $E(\sigma^2_{fsc(X1,X2,...,XP)} \mid \theta)$ is the expectation of the conditional variance of the scoring function, given θ, and $\sigma^2_{fsc(X1,X2,...,XP)}$ is the unconditional variance of the scoring function.

Finally, for those populations in which the BDI-1A is shown to possess satisfactory psychometric properties (i.e., in which it is described by the graded item response model), and possesses adequate reliability, it may be formally tested whether the making of mean comparisons is allowable. Essentially, one tests the hypothesis that the 21 + 63 = 84 parameters of the model are numerically identical in the populations under study. Item parameter invariance is in keeping with the claim that the BDI-1A measures the same thing, in the same way, in the populations under study. Given invariance, one may reasonably assert that the populations differ, at most, with respect to the amount of the construct they manifest (Thissen et al., 1986). That is, at most, they differ with respect to their θ distributions. It is then appropriate to make mean comparisons. The lack of item parameter invariance is evidence for what, in item response theory, is called Differential Item Functioning (DIF) (Thissen et al., 1986).

A number of studies have employed item response theory to characterize the psychometric properties of the BDI-1A. Gibbons, Clarke, VonAmmon-Cavanaugh, and Davis (1985) were the first to publish an item response theory analysis of the BDI-1A. They found that both the unidimensional, two parameter normal ogive and logistic models described their BDI-1A data. They then compared the functioning of the BDI-1A in samples of medically ill patients and depressed inpatients. Their findings suggested the presence of differential item functioning. Unfortunately, as a result of technical limitations, they were forced to dichotomize the BDI-1A items as {0,1} = 0 and {1,2} = 1, thus compromising the integrity of the test. Bouman and Kok (1987), once again dichotomizing the BDI-1A items (this time as {0} = 0 and {1,2,3} = 1), found that the one-parameter logistic model did not describe well the item responses of a sample of depressed patients. They concluded that, within the population of depressed patients, the BDI-1A items did not all measure the same construct, but rather three distinct constructs. Hammond (1995) utilized a one-parameter logistic model (Andrich, 1978) to determine whether differential item functioning was present when comparing a sample of depressed outpatients to a sample of the "general public." On the basis of a DIF index developed by Lord (1980), his findings suggested, for these samples, the presence of DIF. Finally, Santor, Ramsay, and Zuroff (1994) employed nonparametric item response techniques to address the issue of DIF in a comparison of college students and depressed outpatients. They found evidence of DIF and concluded that caution must be exercised in comparing these populations on the basis of the BDI-1A.

The aim of the current work is to further contribute to general knowledge regarding the construct validity of the BDI-1A by assessing whether the BDI-1A behaves satisfactorily within samples drawn from four distinct populations, and then, where appropriate, testing whether population mean comparisons are allowable. One of the populations, the clinically depressed, is the population for which the BDI-1A was derived and in which it was normed. The others, university students from two different universities (Guelph and Western Ontario) and nondepressed psychiatric patients, are populations for which the BDI-1A was not intended for use.

The BDI-II (Beck, Steer, & Brown, 1996) was recently introduced to remedy apparent content validity problems inherent to the BDI-1A. It might then be asked why further effort should be expended on the analysis of the BDI-1A. There exist at least two very good reasons. First, a great number of empirical claims regarding depression have arisen from studies based on scores on the BDI-1A. The validity of these claims is dependent upon the validity of the BDI-1A. In other words, a certain portion of the current knowledge regarding depression is inextricably tied to the construct validity of the BDI-1A. In fact, the investigation of the construct validity of a test never ends. To believe that the introduction of a "new and better" version of a test brings to an end concerns regarding the construct validity of the previous version is a misunderstanding of construct validation theory. The second reason is pragmatic in nature: The BDI-1A is currently more popular than the BDI-II. Since its introduction over four years ago, only 18 studies involving the BDI-II (6 dissertations and 12 articles or chapters) have been published. Furthermore, of the 73 studies involving the BDI published between January and July 2000, 94.5% employed the BDI-1A version of the test.

METHOD

Participants and Measure

The participants were 210 hospitalized depressed inpatients from the Clarke Institute, Toronto, Ontario (150 female, 60 male), 98 nondepressed psychiatric inpatients from the Homewood Institute, Guelph, Ontario (57 female, 41 male), 296 introductory psychology students from the University of Western Ontario (175 female, 121 male), and 328 introductory students from the University of Guelph (203 female, 125 male). The former university is a large law and medicine university located in London, Ontario, while

TABLE 1
Means and Standard Deviations of BDI-1A Items for Each Sample

Item	Depressed In-patients (n = 210)	Nondepressed Psychiatric (n = 98)	Guelph (n = 328)	Western Ontario (n = 296)
1	1.22 (0.92)	0.86 (0.98)	0.57 (0.71)	0.33 (0.57)
2	1.18 (0.93)	0.68 (0.97)	0.38 (0.57)	0.23 (0.45)
3	1.13 (0.96)	0.80 (0.96)	0.29 (0.59)	0.18 (0.43)
4	1.54 (0.84)	1.20 (0.94)	0.57 (0.79)	0.40 (0.69)
5	1.10 (0.90)	0.68 (0.93)	0.36 (0.61)	0.20 (0.49)
6	0.80 (1.13)	0.73 (1.10)	0.27 (0.67)	0.20 (0.57)
7	1.32 (0.85)	0.90 (0.87)	0.59 (0.66)	0.41 (0.61)
8	1.33 (0.85)	1.01 (0.90)	0.63 (0.77)	0.47 (0.64)
9	0.73 (0.81)	0.41 (0.61)	0.33 (0.57)	0.16 (0.42)
10	1.11 (1.07)	1.04 (1.19)	0.47 (0.82)	0.28 (0.69)
11	1.10 (0.77)	1.14 (0.98)	0.64 (0.81)	0.48 (0.79)
12	0.96 (0.85)	0.63 (0.83)	0.27 (0.54)	0.21 (0.46)
13	1.33 (0.94)	1.14 (0.95)	0.36 (0.65)	0.28 (0.61)
14	1.02 (0.98)	0.69 (0.93)	0.55 (0.84)	0.34 (0.69)
15	1.43 (0.72)	1.13 (0.85)	0.66 (0.70)	0.43 (0.66)
16	1.30 (0.99)	1.29 (0.99)	0.60 (0.67)	0.42 (0.67)
17	1.26 (0.82)	1.10 (0.76)	0.85 (0.69)	0.49 (0.60)
18	0.78 (0.88)	0.49 (0.87)	0.38 (0.62)	0.28 (0.61)
19	0.64 (0.93)	0.57 (0.96)	0.24 (0.60)	0.24 (0.57)
20	0.71 (0.78)	0.63 (0.97)	0.34 (0.54)	0.23 (0.47)
21	1.20 (1.07)	1.03 (1.11)	0.20 (0.50)	0.16 (0.49)

the latter is a smaller "comprehensive" university located in Guelph, Ontario. Each participant was administered the BDI-1A (Beck et al., 1979). The BDI-1A is a 21-item test, each item with a 4-option Likert response scale. Option 0 indicates the self-reported absence of a given symptom while option 3 indicates self-reported severe or persistent expression of that symptom. For each item, the respondent is asked to choose the option that best reflects the way he or she has been feeling over the course of the previous week.

Analyses

For each sample, the means and standard deviations of the 21 BDI-1A items were computed. MULTILOG 6(Thissen, 1991) was then employed to fit Samejima's graded model to the item responses. In addition to estimates of the parameters of the graded model, MULTILOG 6 provides the standard asymptotic test of the hypothesis that the model is correct in the population, against the general multinomial alternative. However, this test is not valid if the K^p contingency table of response patterns is sparse, as was the case for each of the four samples of the present study. Three alternative criteria were therefore employed to judge the adequacy of fit of the model to the four samples. First, for each sample, the 84 standardized residuals

$$Z_{j(k)} = (P_{j(k)}-E(P_{j(k)}))/(E(P_{j(k)})*(1-E(P_{j(k)}))/N_{j(k)})^{1/2}$$

were computed. Here, $P_{j(k)}$ is the observed proportion who endorsed option k of item j, $E(P_{ij})$ is the corresponding model implied proportion, and $N_{j(k)}$ is the number of individuals who endorsed option k of item j. If the model is correct, these statistics will, asymptotically, have a standard normal distribution. Hence, if the model is correct, for a given sample, roughly 95% should be less than 1.96 in absolute value. Second, for those samples in which the standardized residuals were judged acceptable, a "pseudo chi-square" goodness of fit index was calculated for each item:

$$\chi^2_j = \Sigma\Sigma N_i[P_{j(k)i}-E(P_{j(k)i})]^2/[E(P_{j(k)i})*(1-E(P_{j(k)i}))]$$

in which $P_{j(k)i}$ is the proportion of those in the interval centred at θ_i who endorse option k of item j, $E(P_{j(k)i})$ is the corresponding model implied proportion, and N_i is the number falling within interval θ_i. These indices quantify the agreement between empirical and model implied option characteristic curves. Adequate agreement was taken as none of the 21 χ^2_j values exceeding two times its degrees of freedom (14 in this case). In all analyses, six θ intervals were employed. Finally, if the fit was deemed acceptable on the basis of the first two criteria, the property of θ-invariance was examined. Essentially, if the graded model describes the BDI-1A in a particular sample, estimates of θ based on subsets of BDI-1A items will be roughly the same in all subsets considered. In particular, if the model provides an adequate fit, then the estimates of θ based on these subsets will be linearly related with a slope, b, close to 1.00, an intercept, a, close to 0, and a Pearson Product Moment Correlation, r_{xy}, close to 1.00 (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). In the present study, θ-invariance was assessed by splitting the BDI-1A items into two sets, one containing odd-, and one, even-numbered items.

For those samples in which the BDI-1A was adequately described by the graded item response model, an appropriate scoring rule was first determined. Specifically, the hypothesis was tested that the 21 slope parameters were equal. A lower bound to reliability was then estimated. Finally, the hypothesis of cross-population item parameter invariance was tested. MULTILOG 6 was used to fit the model simultaneously to these samples while constraining the item parameters to be equal over the samples. The fit of the model under this constraint was assessed using the multifaceted procedure previously described. In the event that the hypothesis of parameter invariance was

TABLE 2
Fit of Graded Item Response Model to Each Sample

Sample	% of Standardized Residuals ≥1.96	% of $\chi^2 > 2$ x df	θ Parameter Invariance Slope	Intercept	r_{xy}
Depressed In-Patients	0	0	0.76	0.00	0.81
Guelph	0	0	0.76	0.04	0.79
Nondepressed Psychiatric	41.7	N/A	N/A	N/A	N/A
Western Ontario	39.3	N/A	N/A	N/A	N/A

TABLE 3
Item Parameter Estimates for Depressed In-patients and Guelph

	Depressed Inpatients				Guelph			
Item	a	τ_1	τ_2	τ_3	a	τ_1	τ_2	τ_3
1	1.52	- 1.39	0.62	1.72	1.89	0.01	1.89	2.74
2	1.82	- 1.14	0.61	1.57	1.87	0.44	2.76	3.14
3	1.41	- 1.01	0.52	1.96	1.85	0.88	2.22	3.16
4	1.44	- 2.50	0.10	1.46	1.51	0.18	2.02	2.49
5	1.17	- 1.20	0.75	2.41	1.46	0.71	2.48	3.61
6	0.87	0.31	1.58	1.99	1.19	1.48	2.75	3.15
7	1.97	- 1.73	0.54	1.42	1.83	- 0.14	2.18	2.81
8	1.25	- 2.14	0.53	1.93	1.37	- 0.05	1.82	2.96
9	1.43	- 0.36	1.53	2.76	1.70	0.69	2.61	3.53
10	0.81	- 1.05	1.45	2.05	1.04	0.80	2.73	2.91
11	1.02	- 1.83	1.37	3.07	1.14	- 0.02	2.36	2.78
12	1.21	- 0.89	0.93	3.07	1.38	1.08	2.84	4.33
13	1.34	- 1.29	-0.15	2.26	1.52	0.81	2.08	3.70
14	0.90	- 0.85	0.87	2.83	0.95	0.52	2.28	3.29
15	1.58	- 2.11	-0.12	2.69	1.46	- 0.20	1.74	4.18
16	0.58	- 2.48	0.97	2.91	0.89	- 0.24	3.54	4.48
17	0.93	- 2.21	0.50	3.11	1.05	- 1.12	2.10	4.04
18	0.68	- 0.50	2.50	3.96	1.02	0.82	3.01	5.41
19	0.32	1.35	4.55	8.43	0.48	3.25	6.44	8.11
20	0.36	- 0.59	0.90	3.34	1.09	0.88	3.38	15.35
21	0.51	-1.59	0.90	3.34	1.37	1.51	2.81	4.87

not supported, the location of the DIF was examined through the computation of an area based DIF measure, one for each of the 84 options, due to Linn, Levine, Hastings, and Wardrop (1981). Essentially, this measure is the average (over the theta continuum from -4 to 4) squared difference between the option characteristic curves of each pair of samples.

RESULTS

Descriptives

The means and standard deviations of the BDI-1A items, for each of the samples, are presented in Table 1.

Fit of the Graded Model

A summary of the fit of the graded model to the BDI-1A item responses for each of the samples is presented in Table 2. For the samples of depressed patients and University of Guelph students, the graded model provided an adequate fit. In the case of both samples, the absolute values of all of the standardized residuals were less than 1.96 and the pseudo-chi square values were all less than 2 times their respective degrees of freedom. In addition, there was evidence of θ-invariance: for the depressed in-patients, b = .76, a = 0, and r_{xy} = 0.81; for the students, b = .76, a = .04, and r_{xy} = .79. For these samples, it was then concluded that the performance of the BDI-1A was in keeping with its theoretical structure. For the samples of nondepressed psychiatric patients and Western Ontario students, the graded model did not provide an adequate fit, 40% of the standardized residuals in each sample larger in absolute value than 1.96. For the samples of

TABLE 4
Estimated DIF for the Response Options of the BDI-1A

Item	Response Option			
	0	1	2	3
1	.06	.07	.02	.04
2	.08	.14	.03	.08
3	.10	.06	.04	.05
4	.18	.15	.03	.03
5	.09	.06	.03	.04
6	.03	.01	.00	.03
7	.09	.13	.02	.07
8	.11	.08	.02	.03
9	.03	.04	.01	.02
10	.06	.03	.00	.02
11	.07	.05	.03	.00
12	.09	.06	.04	.03
13	.11	.05	.09	.05
14	.04	.01	.02	.00
15	.10	.09	.10	.05
16	.07	.07	.02	.03
17	.02	.04	.03	.01
18	.03	.02	.01	.02
19	.03	.01	.01	.00
20	.06	.07	.03	.09
21	.12	.04	.03	.05

Note. Mean DIF = .05

depressed in-patients and Guelph students, parameter estimates are presented in Table 3, and for items 2, 4, 7, 19, and 21, option characteristic curves in Appendices A and B, respectively.

For both groups, the curves for item 19, "concentration difficulty," are relatively flat, indicating that this item is not strongly related to the latent dimension, and, hence, is poorly functioning. All of the other items appear to be functioning satisfactorily.

Scoring Rules and Reliability

For both the depressed in-patients ($\chi^2(20) = 137.9$, $p < .05$) and Guelph students ($\chi^2(20) = 98.6$, $p < .05$), the hypothesis of equal item slope parameters was rejected. Hence, it was concluded that the appropriate scoring rule for each sample was the weighted composite $\Sigma a_j X_j$. The estimated lower bound to the reliability of this composite was .88 for both samples. All told, it was concluded for both of these groups that the psychometric performance of the BDI-1A was satisfactory.

Test of Parameter Invariance

For the depressed in-patients and Guelph students, the hypothesis of cross-population parameter invariance was rejected. Under the constraint of parameter invariance, the graded model fit neither group adequately, with 10.7% of the standardized residuals exceeding 1.96 in absolute value for the in-patients, and 54.8% for the students. It was therefore concluded that, with respect to these groups, differential item functioning was present and mean comparisons not allowable.

With respect to the depressed in-patients and Guelph students, Table 4 presents DIF measures for the 84 options of the BDI-1A. As noted by Santor et al. (1994), there do not exist intuitively pleasing benchmarks by which to compare DIF values such as these. It is clear, however, that DIF is present to greatest degree in options 0 ("I get as much pleasure as I ever did from the things I enjoy") and 1 ("I don't enjoy things as much as I used to") of item 4, option 1 of items 2 ("I feel more discouraged about my future than I used to be") and 7 ("I have lost confidence in myself"), and option 0 of items 21 ("I have not noticed any recent change in my interest in sex") and 13 ("I make decisions about as well as ever"). Plots of these response characteristic curves are presented in Appendices A and B. The Guelph students have a higher probability of endorsement of option 0 of item 4 than the depressed in-patients, at all points on the theta continuum, while the opposite is true for option 1 of item 4 (plots for the other items are available by e-mailing the corresponding author).

DISCUSSION

The current analysis is one of very few to employ a psychometrically sound framework to assess construct validity issues pertaining to the BDI-1A. The results are suggestive of the difficult task a researcher faces if he or she desires the sanction of psychometrics to (1) employ a simple, unweighted composite of BDI-1A items (i.e., total score) as a measure of depression; 2) employ the BDI-1A in populations other than that in which it was normed; or 3) use the BDI-1A total score as input into tests of hypotheses concerning population mean differences in level of depression. In particular, the BDI-1A was found to be employable in only two of the four samples considered (depressed in-patients and university students from the University of Guelph), and, within these samples, could not justifiably be scored as a simple total score. The estimated lower bounds to reliability were, it should be noted, quite acceptable for both of these groups. Finally, the BDI-1A weighted total scores, "depression scores," derived for the depressed in-patients and the University of Guelph students were found not to be comparable, and hence could not be used as input into the making of mean comparisons with respect to level of depression. DIF was especially noticeable in items 2, 4, 7, 13, and 21. Item 19, on the other hand, was found to be poorly functioning within both groups.

Various interpretations of these results are possible. Each issue is considered in turn.

Issue 1

It is true that, in many instances, the flouting of "fancy" scoring rules in favour of a simple unit-weighted scoring rule will "make little difference" to results involving the BDI-1A. However, whether this is true is a case-by-case consideration. For example, although the correlations of a composite of positively correlated BDI-1A items with external criteria of interest may be relatively invariant with respect to scoring rule, the information function of the test may well not be. Until, in a particular context, choice of scoring rule is shown to make little difference, the optimal scoring rule should be preferred. In fact, Santor et al. (1994) present examples in which gains in measurement precision are made by applying nonlinear transformations to the option weights of the BDI-1A.

Issue 2

The test was found to have an empirical structure in keeping with its theoretical structure in both the samples of depressed in-patients and the Guelph students. This means essentially that, in both samples, one can produce a composite, or total score, based on the test that reflects a single dimension of variation, whatever be this dimension. Unidimensionality is the psychometric justification for the practice of producing such single composites to represent a test. Since the BDI-1A was designed to measure the severity of depression of depressed individuals, and since the test was normed in populations of depressed individuals, it is tempting to conclude, given this finding, that the test does, at least for the depressed in-patients, measure severity of depression. Psychometrics, however, provides no licence for such a conclusion. It may only be concluded that the test performed in a manner that is in keeping with its theoretical structure. Furthermore, it is entirely possible (and in fact likely, given the lack of parameter invariance) that this measured dimension is different in the two samples.

For both the University of Western Ontario and nondepressed psychiatric samples, the test was found not to perform psychometrically as its theoretical structure implies that it should. This means that for these groups there exists no psychometric justification for the production of a BDI-1A composite score. Hence, the present study provides no justification for the use of the test in these settings.

Some might feel perplexed by the differences in the psychometric properties of the test in the two university samples. Generally speaking, the desire for apparent consistency in results is naive. A set of test scores is the product of a complex interaction of test and subject population characteristics (Bejar, 1983). And, of course, population characteristics are difficult to define, yet alone control for. It will rarely be possible to construct scientifically rigorous accounts of the reasons for cross-population differences in test structure. In any case, the identification of the "reason" for the differences between these samples is not, at least initially, as important as the discovery that there were differences, a fact which undermines any attempts to make the usual claim that the test measures the same thing in the these two contexts.

Issue 3

Even though for both the Guelph students and the depressed in-patients, psychometric justification was found to exist for the production of a BDI-1A composite score, this does not mean that this score is a measure of the same construct. In the current work, the hypothesis of parameter invariance was not supported, a finding that is usually taken to mean that at least some of the items are measuring differentially in the two groups. The items that manifest the greatest degree of DIF were items 4, 2, 7, 13, and 21. A number of researchers (e.g., Chrisjohn & Bradley, 1989; Gotlib, 1984; Hammond, 1995; Tanaka-Matsumi & Kameoka, 1986) have in fact suggested that, in student populations, the BDI-1A measures not depression, but, instead, general psychopathology or anxiety. In the current work, scores on anxiety and psychopathology measures were not available. Hence, the reasonableness of this particular explanation could not be assessed.

It should, in general, be anticipated by the psychological scientist that cross-population parameter invariance, a desirable outcome from a test theory perspective, will, in practice, be an exceedingly rare occurrence. Hanna (1984), for example, documented a case in which invariance of item structure did not exist when the same group of children were measured on the same two tests, two years apart. What should the user of tests in general, and of the BDI-1A in particular, take from this fact? Just that, in general, mean comparisons between distinct populations made on the basis of the BDI-1A will seldom be justifiable. Once again, psychometric issues are tied inextricably to the populations of individuals the psychologist wishes to measure. Hence, in practice these issues will be no less complex than these very individuals.

In conclusion, it might be objected that the psychometric requirements placed in this study on the BDI-1A were too strict. But this is a moot point: Psychometric justification of test use is required by the American Psychological Association, and the requirements described herein are those that justify practices (1), (2), and (3). There seems to us to exist a dangerous tension between psychometric requirements and the psychologist's desire to get out and "do research." The

former should never be given short shrift, lest the latter be fatally compromised.

This research was supported in part by a SSHRC Small Grant awarded to the second author.

Correspondence concerning this article should be addressed to Michael D. Maraun, Department of Psychology, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada (Fax: (604) 291-3427; E-mail: maraun@sfu.ca).

References

American Psychological Association, American Research Association & National Council on Measurement in Education (1985). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43* (4), 561-573.

Beck, A.T. (1972). *Depression: Causes and treatments*. Philadelphia, PA: University of Pennslyvania Press.

Beck, A.T., Rush, A.J., Shaw, B.F., & Emery, G. (1979). *Cognitive therapy of depression*. New York: Guildford Press.

Beck, A.T., Steer, R.A., & Brown, G.K. (1996). *Beck Depression Inventory (2nd ed.)*. San Antonio, TX: Harcourt Brace & Company.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561-571.

Bejar, I.I. (1983). Introduction to item response models and their assumptions. In R.K. Hambleton (Ed.), *Applications of Item Response Theory*. Vancouver, BC: Educational Institute of British Columbia.

Bouman, T.K., & Kok, A.R. (1987). Homogeneity of Beck's Depression Inventory (BDI-1A): Applying Rasch analysis in conceptual exploration. *Acta Psychiatrica Scandinavica, 76* (5), 568-573.

Chrisjohn, R.D., & Bradley, H. (1989, June). *The relation between anxiety and depression: A facet approach*. Paper presented at the 50th annual meeting of the Canadian Psychological Association, Halifax, NS.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52* (4), 281-302.

Gibbons, R.D., Clarke, D.C., VonAmmon Cavanaugh, S., & Davis, J.M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research, 19* (1), 43-55.

Golin, S., & Hartz, M.A. (1979). A factor analysis of the Beck Depression Inventory in a mildly depressed population. *Journal of Clinical Psychology, 35* (2), 322-325.

Gotlib, I.H. (1984). Depression and general psychopathology in university students. *Journal of Abnormal Psychology, 93* (1), 19-30.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* London: Sage.

Hammond, S.M. (1995). An IRT investigation of the validity of non-patient analogue research using the Beck Depression Inventory. *European Journal of Psychological Assessment, 11*(1), 14-20.

Hanna, G. (1984). The use of a factor analytic model for assessing the validity of group comparisons. *Journal of Educational Measurement, 21* (2), 191-199.

Hill, A., Kemp-Wheeler, S., & Jones, S. (1986). What does the Beck Depression Inventory measure in students? *Personality and Individual Differences, 7* (1), 39-47.

Linn, R., Levine, M., Hastings, C., & Wardrop, J. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5* (2), 159-173.

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

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

Maraun, M., Jackson, J.S., Luccock, C.R., Belfer, S.E., & Crisjohn, R.D. (1998). CA and SPOD for the analysis of tests comprised of binary items. *Journal of Educational and Psychological Measurement, 58* (6): 916-928.

Piotrowski, C., & Keller, J. (1992). Psychological testing in applied settings: A literature review from 1982-1992. *Journal of Training and Practice in Professional Psychology, 6* (2), 74-84.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs, 34* (17, pt. 2).

Santor, D.A., Ramsay, J.O., & Zuroff, D. (1994). Nonparametric item analysis of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6* (3), 255-270.

Tanaka-Matsumi, J., & Kameoka, V. (1986). Reliabilities and concurrent validations of popular self-report measures of depression, anxiety and social desirability. *Journal of Consulting and Clinical Psychology, 54* (3), 328-333.

Thissen, D. (1991). *MULTILOG users guide: Multiple, categorical item analyses and test scoring using item response theory* (version 6) [computer program]. Chicago, IL: Scientific Software.

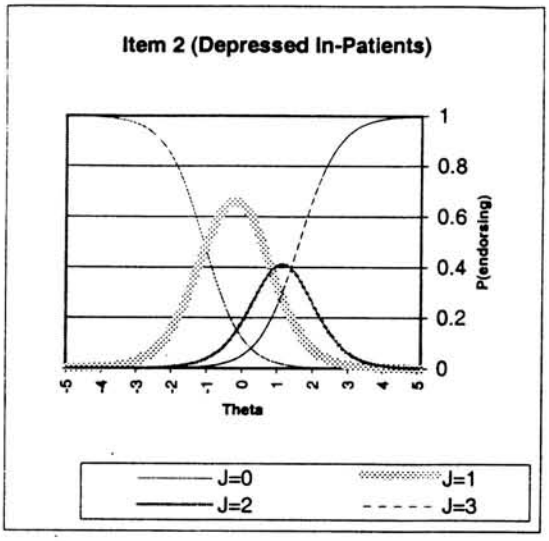
Thissen. D., Steinberg, L., & Gerrard (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99* (1), 118-128.

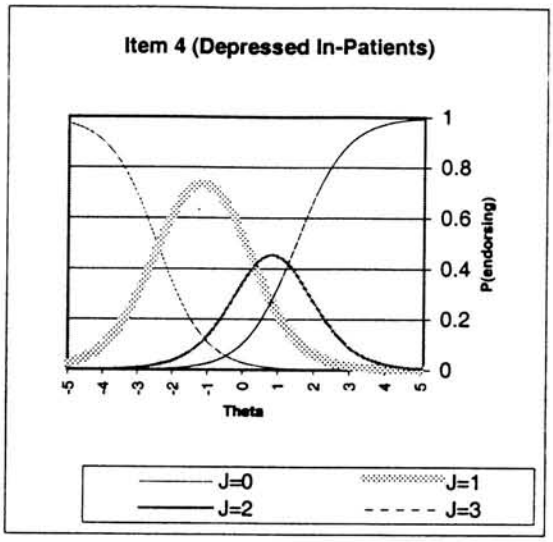
Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J.

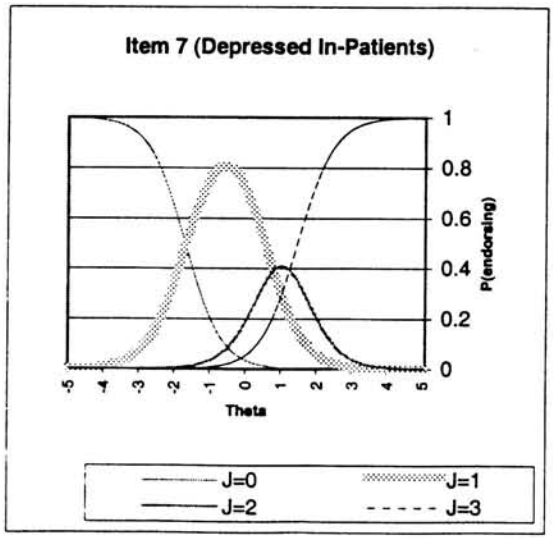
(1983). An item responsetheory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement, 7*(2), 211-226.

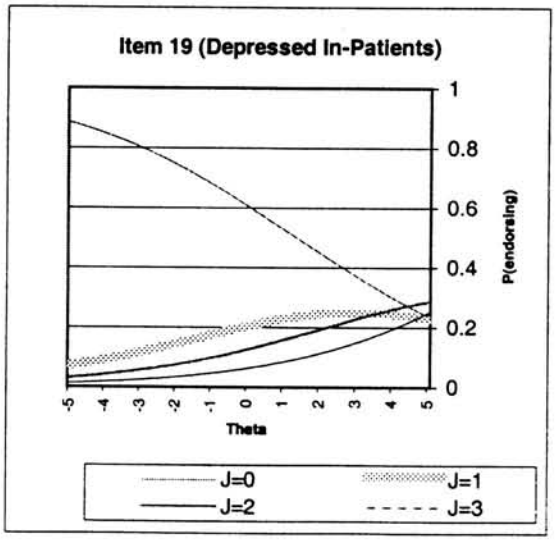
Received January 7, 2000
Revised January 31, 2001
Accepted February 15, 2001

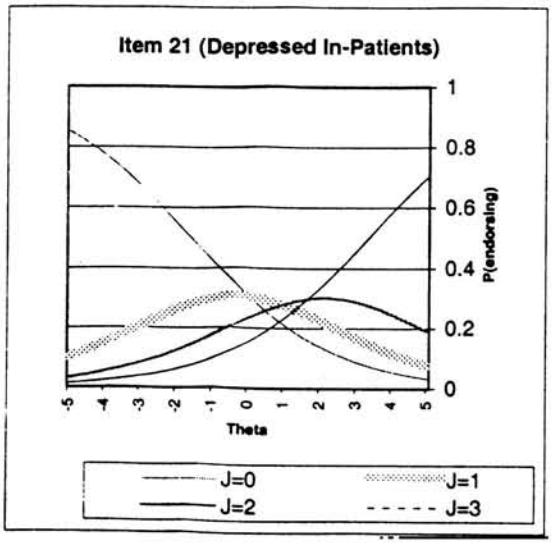
Appendix A


Item 2 (Depressed In-Patients)
P(endorsing)
Theta
J=0
J=1
J=2
J=3

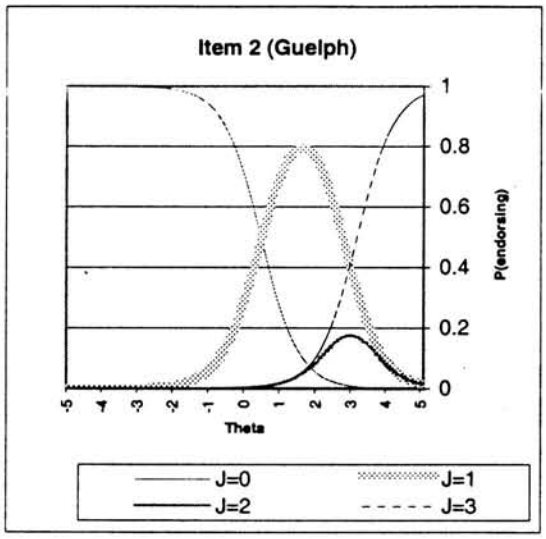

Item 4 (Depressed In-Patients)
P(endorsing)
Theta
J=0
J=1
J=2
J=3

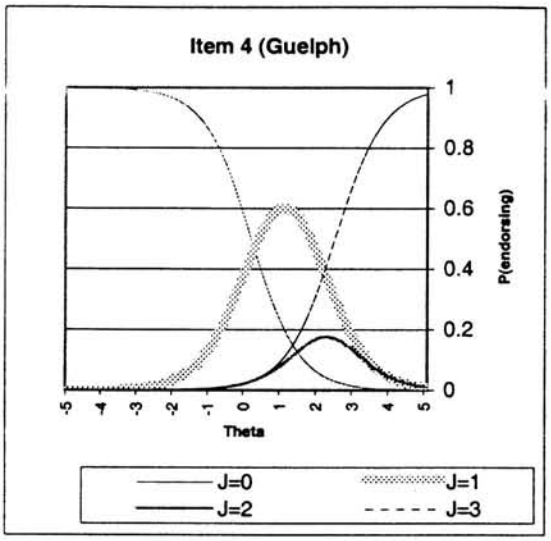

Item 7 (Depressed In-Patients)
P(endorsing)
Theta
J=0
J=1
J=2
J=3

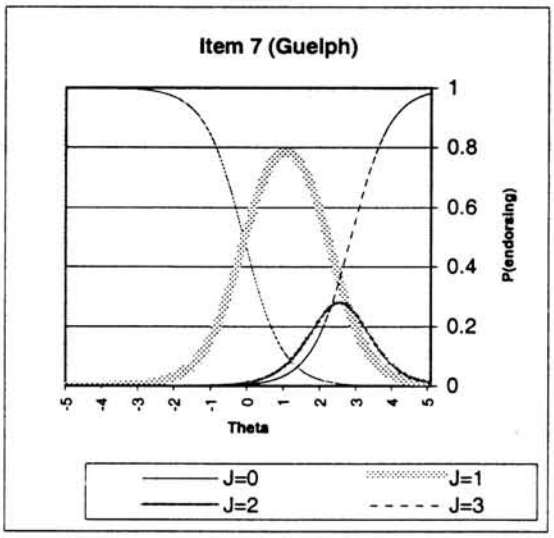

Item 19 (Depressed In-Patients)
P(endorsing)
Theta
J=0
J=1
J=2
J=3

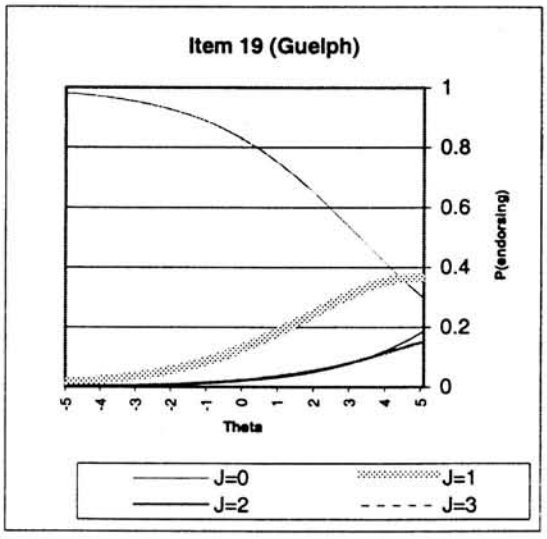

Item 21 (Depressed In-Patients)
P(endorsing)
Theta
J=0
J=1
J=2
J=3

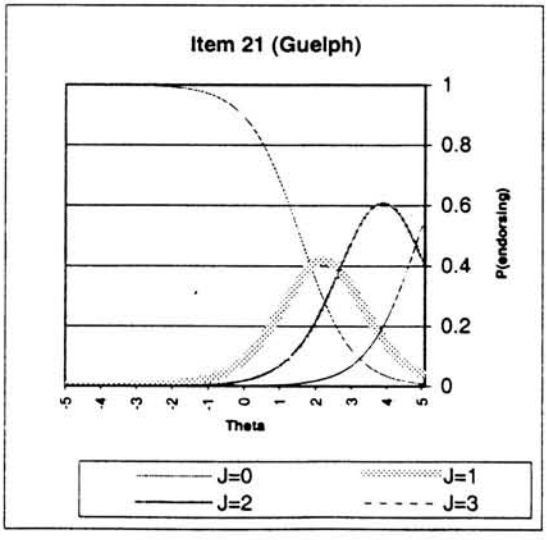
Appendix B


Item 2 (Guelph)
P(endorsing)
Theta
J=0
J=1
J=2
J=3


Item 4 (Guelph)
P(endorsing)
Theta
J=0
J=1
J=2
J=3


Item 7 (Guelph)
P(endorsing)
Theta
J=0
J=1
J=2
J=3


Item 19 (Guelph)
P(endorsing)
Theta
J=0
J=1
J=2
J=3


Item 21 (Guelph)
P(endorsing)
Theta
J=0
J=1
J=2
J=3