AMR Applied Multivariate Research

Volume 13
Issue 2

In this issue:

Journal of the Society of Applied Multivariate Research

Applied Multivariate Research, Volume 13, No. 2, 2010, 113-128

THE EFFECTS OF ESTIMATOR CHOICE AND WEIGHTING STRATEGIES ON CONFIRMATORY FACTOR ANALYSIS WITH STRATIFIED SAMPLES

Bradley J. Brummel[1]
The University of Tulsa

Fritz Drasgow
University of Illinois at Urbana-Champaign

ABSTRACT

Survey researchers often design stratified sampling strategies to target specific subpopulations within the larger population. This stratification can influence the population parameter estimates from these samples because they are not simple random samples of the population. There are three typical estimation options that account for the effects of this stratification in latent variable models: unweighted maximum likelihood, weighted maximum likelihood, and pseudo-maximum likelihood estimation. This paper examines the effects of these procedures on parameter estimates, standard errors, and fit statistics in Lisrel 8.7 (Jöreskog & Sörbom, 2004) and Mplus 3.0 (Muthén & Muthén, 2004). Options using several estimation methods will be compared to pseudo-maximum likelihood estimation. Results indicated the choice of estimation technique does not have a substantial effect on confirmatory factor analysis parameter estimates in large samples. However, standard errors of those parameter estimates and RMSEA values for assessing of model fit can be substantially affected by estimation technique.

Keywords: stratified samples, pseudo-maximum likelihood, parameter estimation, weighting strategies, latent variable modeling

INTRODUCTION

Researchers often employ complex sampling methodologies when surveying large populations. These methodologies include stratified sampling, cluster sampling, over-sampling specific subpopulations, and other designs that lead to unequal probabilities of selection within the population (Skinner, Holt, & Smith, 1989). For example, the Defense Manpower Data Center stratified their survey sample on the basis of reserve component, reserve program, gender, paygrade group, racial/ethnic group membership, and activation status for the 2004 Workplace and Gender Relations Survey of Reserve Component Members

[1] Correspondence and requests for reprints should be addressed to Bradley J. Brummel, Department of Psychology, Lorton Hall, The University of Tulsa, 800 S. Tucker Dr., Tulsa, OK, 74104 or email at bradley-brummel@utulsa.edu.

(Kroeger, 2004). These methodologies are used to obtain more reliable parameter estimates of subpopulations within the survey population. Stratified sampling can also be used to ensure a more closely matched sample to the actual population by taking random samples of a proportional size to the population within each subpopulation. Subpopulations are often chosen based upon geographic area, age, or minority group status within the population of interest. When the sampling procedure does not result in a simple random sample of the population, the survey sample will violate a basic assumption of most inferential statistical analyses. This violation can lead to biased estimates of population parameters unless the analysis accounts for the survey design (Skinner, et al.). Corrections are possible when informative sampling weights are available; in this situation, the sample cases can be weighted to appropriately represent the population (Asparouhov, 2005).

Correcting the sample estimates of means and variations to overall population parameter estimates is a straightforward task when a researcher has informative sampling weights. This is done by weighting each case in the sample back to its frequency in the overall population and adjusting the parameter estimates by this frequency. The population mean can be estimated by the weighted mean

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i},$$

where w_i is the weight for the i^{th} observation and y_i is the observed score for the i^{th} observation (Stapleton, 2002). The sampling variance of the estimator can be calculated by

$$\hat{\text{var}}(\hat{\mu}) = \frac{\sum_{i=1}^{n} w_i (y_i - \hat{\mu})^2}{\sum_{i=1}^{n} w_i \left(\sum_{i=1}^{n} w_i - 1 \right)},$$

where w_i is the weight for the i^{th} observation and y_i is the observed score for the i^{th} observation (Stapleton).

Obtaining accurate population estimates of parameters and fit statistics for latent variable models parameters is not as straightforward. These estimates require appropriate weighting throughout the process of matrix operations required for estimates of the model parameters and fit statistics. This issue has gained research interest in the past few years resulting in options for incorporating informative sampling weights in latent variable modeling programs including Lisrel Version 8.7 and Mplus Version 3.

Researchers have typically used one of three main weighting and estimation techniques for estimating population parameters for latent variables from sample data with unequal probability of selection. These methods are unweighted maximum likelihood (UML), weighted maximum likelihood (WML), and, pseudo-maximum likelihood (PML) estimation techniques. There has also been research examining the best scheme for scaling the sampling weights for the WML estimation (Kaplan & Ferguson, 1999, Stapleton, 2002). These procedures include using the raw sample weights, relative sample weights, or effective

sample weights. Each of these estimation techniques has implications for estimation bias and the resulting fit statistics for the estimated models.

This paper will describe these methods for estimating models from complex sampling designs. Then, it will examine the resulting parameter estimates, standard errors, and fit statistics for each of these techniques in the context of confirmatory factor analysis (CFA) of a measure of organizational commitment in a stratified sample using the procedures available in Lisrel 8.7 and Mplus 3.0. Finally, we will make suggestions regarding the utility of incorporating the various methods in typical scale evaluation and development, considering sample size and the degree of stratification.

Estimation methods for data from stratified samples

The three typical estimation methods for incorporating the information provided by knowledge of the sampling design into the estimation of population parameters in latent variable models are UML, WML, and PML. These methods also apply to the more limited use of sampling weights in the estimation of CFA parameters from stratified samples. By examining the difference between these techniques it is possible to make an informed decision about whether it is valuable to use the more complicated methods in a specific application.

Unweighted maximum likelihood. The UML estimation technique is not actually a technique for using the sampling weights from a study design. It is the practice of ignoring the information provided by the sampling weights. The sample data are treated as if they were a simple random sample of the larger population. The population parameters are also estimated as if the sample accurately reflected the distribution of individuals within the population of interest (Asparouhov, 2005).

For the estimation of some parameters, such as the population mean, an unweighted analysis can provide grossly biased estimates. For example, suppose Subpopulation A contains 90% of the cases in a population and has a mean of $\mu_A = 10$ and Subpopulation B contains the remaining 10% of the population and has a mean of 0. If a researcher draws half of his/her sample from each population, the sample mean will converge to

$$.5\ (10) + .5\ (0) = 5,$$

rather than the true population mean

$$.9\ (10) + .1\ (0) = 9.$$

Thus it is clear that this method will provide inaccurate parameter estimates when one or more of the subpopulations differ substantially on certain levels of a measured variable or have different interrelationships between some of the variables or items. Using UML estimation methods also tend to downwardly bias the standard error estimates from stratified samples (Kaplan & Ferguson, 1999).

UML estimation has to be used if information about the sampling design is unavailable or non-informative. If there are varying response rates by subgroup or missing demographic information, then there may be little or no information available to create accurate sampling weights. If the sample is quite large in proportion to the population of interest, then there may be little to be gained by using the sample weights in the latent variable model estimation even if the sample weights are informative. UML is the simplest way to handle a stratified sample in latent variable model estimation, and this method can be implemented with any latent variable modeling statistics package.

Weighted maximum likelihood. The WML estimation technique uses the weighted correlation matrix to estimate model parameters. However, there are multiple schemes that can be used for computing the weighted correlation matrix. Three of these are raw sample weights (WML), relative sample weights (WRML), and effective sample weights (WEML) (Stapleton, 2002). Relative and effective sample weights are suggested to normalize the weighting scales (Potthoff, Woodbury, & Manton, 1992). Raw sample weights are the weights that produce the population size when applied to the sample. When these weights are applied, the effective sample N equals the population N. This method allows unbiased estimation of the population parameters but produces standard errors and confidence intervals that are too small (Asparouhov, 2005). This results in overly liberal statistical inferences when using the chi-square difference tests of model fit.

Relative and effective sample weights (Potthoff, et al., 1992) normalize the weights by scaling the raw sample weights. Relative weights are calculated by multiplying the raw weights by

$$\frac{n}{\sum_{i=1}^{n} w_i}$$

where w_i is raw weight for the i^{th} observation. These relative sample weights recreate the actual sample size when applied to the sample because the average weight is equal to 1. Interestingly, relative sample weights have also been shown to produce downwardly biased estimates of sampling variance in simulation studies (Stapleton, 2002).

Effective weights are calculated by multiplying the raw weights by

$$\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i^{\,2}}.$$

These effective sample weights are designed to recreate the sample size that would produce an equivalent amount of information had the sampling procedure been a simple random sample. Therefore, the *n* for the effective sample weighting technique is always less than or equal to the actual sample and relative sampling technique *n* (Stapleton, 2002). This weighting technique has been shown to produce approximately unbiased estimates of model parameters and their sampling variances in simulation studies (Stapleton).

Mplus can approximate the raw sample weighted WML technique by treating the sample weights as if they were integer frequency weights (Asparouhov, 2005). It can not approximate the other weighting options as most weights in these designs will be between zero and one and will lose large amounts of information if rounded to the nearest integer (Muthén, 2005). Lisrel version 8.7 can incorporate any of these options for sample weights into its latent variable estimating procedures. This can be done through either the weight cases option or the survey design feature in Prelis. However, these weighting schemes do not remove all bias in the estimates. To accurately estimate all of the relevant fit statistics, parameters, and standard errors, the PML estimation technique should be used (Pfefferman, et al., 1998; Asparouhov).

Pseudo-maximum likelihood. The PML estimation technique maximizes the weighted log-likelihood function,

$$Log(L) = \sum_i w_i \log(L_i),$$

where L_i is the likelihood of the i^{th} observation. This method calculates the covariance matrix by

$$\left(\left(\log(L)\right)'' \right)^{-1} \left(\sum_i w_i^2 \left(\left(\log(L_i)\right)' \right)^T \left(\log(L_i)\right)' \right) \left(\left(\log(L)\right)'' \right)^{-1},$$

where ´ and ´´ represent the first and second derivative, *T* indicates the transpose of a matrix, and the sum is over all individuals in the sample. PML provides consistent estimates under all sampling strategies (Asparouhov, 2005). This estimation method provides consistent parameter estimates for stratified samples while maintaining the accuracy of chi-square tests for model fit (Pfefferman, et al., 1998, Skinner 1989). PML is currently available in the statistics package Mplus Version 3.0. PML has been found to provide accurate fit statistics, parameter estimates, and estimated standard errors regardless of the scheme for the sample weights (Pfefferman, et al.).

Study design

This study will examine the parameter estimates, standard errors, and fit statistics obtained from the UML, WML, and PML techniques for a CFA model of an organizational commitment scale (Gade, Tiggle, & Schumm, 2003) included in the 2004 Workplace and Gender Relations Survey of Reserve Component Members using the options available in Mplus 3.0 and Lisrel 8.7. These estimation techniques will be applied to a large, stratified sample with informative sampling weights and to small subsets of that sample to examine the effects of sample size. The results from the UML and various WML strategies will be compared to the PML technique. These analyses will allow us to make suggestions regarding the utility of the various methods for typical scale evaluation and development, considering sample size for the degree of stratification in this sample.

METHOD

Participants

The 2004 Workplace and Gender Relations Survey of Reserve Component Members was distributed to a stratified random sample of 76,031 Reserve component members from March 19, 2004 to June 21, 2004 using paper and pencil and web-based formats (Riemer, 2004). This yielded a usable sample of 26,443 United States Armed Forces reservists (men, *n* = 12,902, 49%, women, *n* = 13,541, 51%).

Data were weighted to reflect the Reserve component population as of March 2004. The final sample weights were calculated using a three-step process. First the base weights were calculated to reflect variable probabilities of selection. The stratification categories included Reserve component, Reserve program, gender, paygrade group, racial/ethnic group membership, and activation status (Kroeger, 2004). The second step adjusted the base weights for non-response. Finally, the non-response-adjusted weights were adjusted to the known

population totals as of the start of data collection. Complete details of weighting and response rates are reported by Flores-Cervantes, Jones, and Wilson (2004).

Measure

Meyer and Allen's (1990) Organizational Commitment scale was modified for more efficient use within military samples (Table 1). This modification involved removing items assessing normative commitment, using a military referent, and phrasing all items positively. It includes 4 items that assess affective commitment and 4 items that assess continuance commitment (Gade, et al., 2003). The scales demonstrated adequate reliability (alpha = 0.91 for affective commitment and 0.88 for continuance commitment) and fit as a two factor model of commitment in the 2004 Workplace and Gender Relations Survey of Reserve Component Members sample (Ormerod, et al., 2005). This measure was analyzed with two underlying factors (Figure 1). Scale items were constrained to load only on the theoretically appropriate factor.

Table 1
Organizational Commitment scale (Gade, Tiggle, & Schumm, 2003)

Affective Commitment (AC)	
(x_1)	I feel like "part of the family" in the military.
(x_2)	The military has a great deal of personal meaning for me.
(x_3)	I feel a strong sense of belonging to the military.
(x_4)	I feel emotionally attached to the military.
Continuance Commitment (CC)	
(x_5)	It would be too costly for me to leave the military in the near future.
(x_6)	I am afraid of what might happen if I quit the military.
(x_7)	Too much in my life would be interrupted if I decided I wanted to leave the military.
(x_8)	One of the problems of leaving the military would be the lack of available alternatives.

Note: Items response scale was a five point scale from *strongly agree* to *strongly disagree*

Analytic Strategies

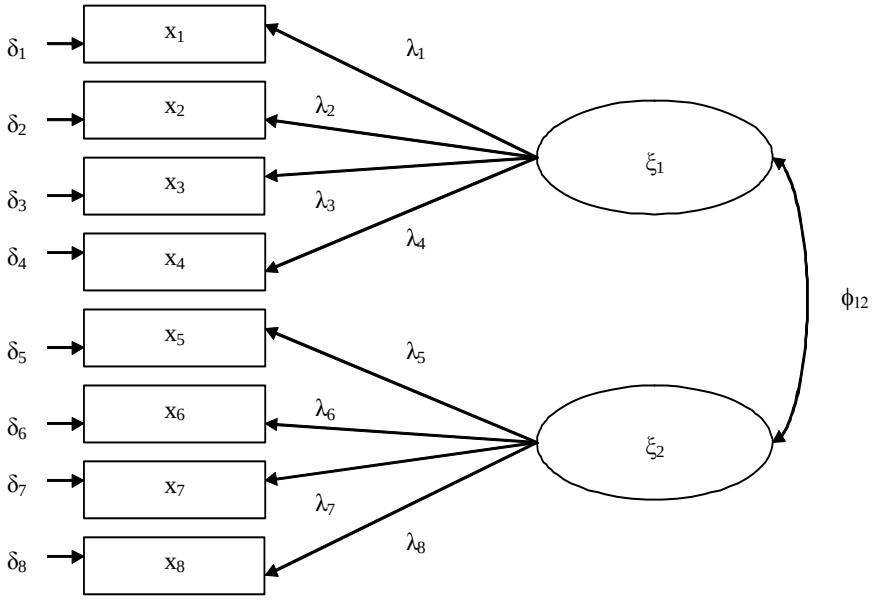
Two-factor CFAs of the 8 item measure of affective and continuance commitment were conducted using the UML, WML, and PML estimation methods in Mplus 3.0 and using the UML, WML, WRML, and WEML methods in Lisrel 8.7. The WRML and WEML methods could not be used in Mplus because the program requires integer weights to conduct WML estimation with the UML estimator (Muthén, 2005). The WML raw weights were rounded to the nearest integer in the Mplus WML analysis so that the results of using sampling weights as frequency rates could be investigated (Asparouhov, 2005). The UML estimation method is identical in both software programs and therefore will only be reported once in the rest of the paper.

Complete data on the commitment measure were available for 98.1% (25,947) of the sample. Due to the high percentage of participants without missing data, participants with any missing data on the measure were deleted from further analyses. This resulted in the removal of 496 participants (1.9%). The analyses were conducted using this sample of participants with complete data for the measure and for 10 randomly selected subsets of 250 participants

from the larger sample. The average of the results for the 10 smaller subsets has been reported.

The informative sample weights in the 25,947 person group ranged from 0.96 to 183.70 with a mean of 29.70 and a median of 13.68 (Figure 2). These weights resulted in an overall population size of 770,641 when applied to the full sample. The relative sample weights ranged from 0.03 to 6.20 with a mean of 1.00 and a median of 0.46. The effective sample weights ranged from 0.01 to 2.85 with a mean of 0.46 and a median of 0.21. In the 250 participant subsets, the informative sample weights were rescaled by a factor of 103.78 (25,946/250) to correspond to weights needed to recreate the population. The relative and effective sample weights for each of the 250 subsets were calculated as described above.

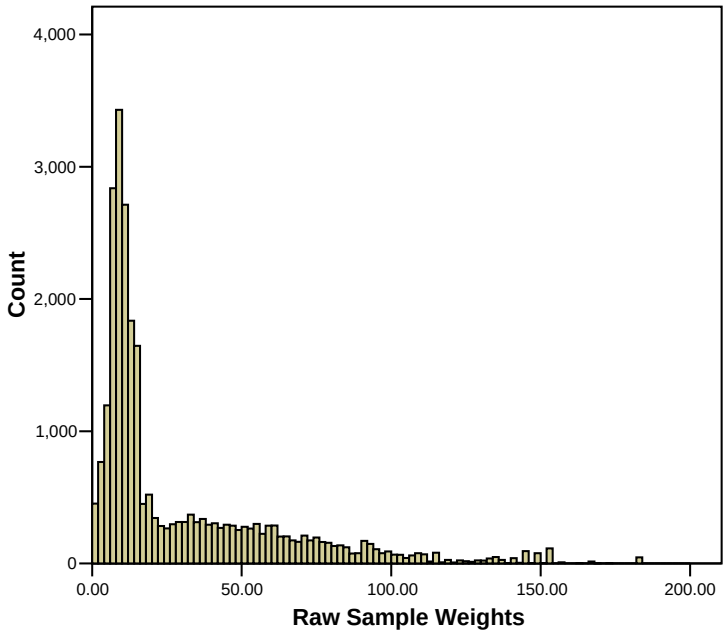
Figure 1
Two factor CFA of Affective and Continuance Commitment measure



RESULTS

The estimates of means, variances, factor loadings, and factor correlation are given in Table 2. The largest differences in parameter estimates were observed for the item means and variances for the unweighted (UML) versus the weighted (WML, WRML, WEML, and PML) techniques. The weighted maximum likelihood estimation methods and the pseudo-maximum likelihood estimation method produced identical parameter estimates; however, the standard errors associated with these estimators were different.

Figure 2
Distribution of raw sampling weights



The use of the WML estimator in Mplus (WML(M)), which treats sample weights as frequency weights, resulted in much smaller standard error estimates than the PML estimator (Table 3); however, the use of WML techniques in Lisrel (WML(L)) resulted in standard error estimates identical to PML. This was also true for all weighting methods as the WML, WRML, and WEML methods all produced the same standard errors estimates. This occurs because Lisrel uses the actual sample size rather than the population size, relative sample size or effective sample size when estimating standard errors, so the weighting scheme does not affect these estimates. The UML standard errors were also smaller than the PML standard error estimates.

The resulting chi-square values, RMSEA, and effective sample sizes for the estimation techniques are presented in Table 4. The only fit statistics available from the WML methods in Lisrel were the full information UML Chi-square and the RMSEA because the program assumes that there is missing data when the weights are applied in Prelis. Also, the inclusion of sample weights in the PML technique does not allow chi-square difference tests to be used because the chi-square values are not distributed as chi-square (Muthén & Muthén, 2004).

The WML(M) technique produced a much larger Chi-square value than the other techniques due to the inferred sample size. The UML Chi-square was also larger than the values for the other WML techniques and the PML technique. The WML weighting schemes again produced equivalent values in Lisrel, but not the same as the PML technique.

Table 2
Parameter estimates for the full sample (N=25,947) using UML, WML, and PML estimators

	Unstandardized			Standardized		
Statistic	**UML**	**WML****	**PML**	**UML**	**WML****	**PML**
μ_1	3.554	3.570	3.570			
μ_2	3.628	3.612	3.612			
μ_3	3.513	3.506	3.506			
μ_4	3.237	3.205	3.205			
μ_5	2.925	2.886	2.886			
μ_6	2.578	2.551	2.551			
μ_7	2.529	2.496	2.496			
μ_8	2.641	2.624	2.624			
λ_1	1.000*	1.000*	1.000*	.887	.877	.876
λ_2	1.055	1.024	1.024	.892	.897	.897
λ_3	1.155	1.187	1.187	1.048	1.040	1.040
λ_4	1.059	1.127	1.127	.996	.988	.988
λ_5	1.000*	1.000*	1.000*	.918	.940	.940
λ_6	1.155	1,154	1,154	1.095	1.085	1.085
λ_7	1.187	1.188	1.188	1.120	1.117	1.117
λ_8	.949	.932	.932	.858	.877	.877
δ_1	.362	.430	.430			
δ_2	.290	.320	.320			
δ_3	.150	.200	.200			
δ_4	.285	.415	.415			
δ_5	.439	.653	.653			
δ_6	.251	.403	.403			
δ_7	.209	.324	.324			
δ_8	.495	.694	.694			
ϕ_{12}	.233	.349	.349	.389	.424	.424
ϕ_{11}	.638	.768	.768	1.000	1.000	1.000
ϕ_{22}	.561	.884	.884	1.000	1.000	1.000

Note: *Fixed value used for identification; **WRML and WEML methods provided identical parameter estimates to WML

The various weighting strategies for the WML methods produced greatly different RMSEA values for the model. The UML and WML(M) RMSEA values were substantially larger than the PML value. It is important to note that these values would result in different conclusions about the fit of the model when using standard criteria such as RMSEA < .05 constituting reasonably good fit (Browne & Cudeck, 1993): The UML, WML(M), and WEML solutions would be judged as providing poor fits, the WRML and the PML solutions would be evaluated as satisfactory, and the WML(L) solution might be viewed as aberrant because the uncommonly good fit of the model. It is important to note that there is disagreement as to the

standards for model fit, especially for applied research. Whichever standards are chosen, the varying RMSEA values do create the possibility of reaching different conclusions regarding the acceptability of the model fit.

Table 3
Standard Error estimates for the full sample (N=25,947) using UML, WML, and PML estimators

Statistic	UML	WML(M)*	WML(L)**	PML
μ_1	.007	.001	.010	.010
μ_2	.007	.001	.010	.010
μ_3	.007	.001	.010	.010
μ_4	.007	.001	.011	.011
μ_5	.008	.001	.011	.011
μ_6	.008	.001	.011	.011
μ_7	.008	.001	.011	.011
μ_8	.007	.001	.011	.011
λ_1	------	------	------	------
λ_2	.007	.001	.010	.010
λ_3	.007	.001	.011	.011
λ_4	.007	.001	.012	.012
λ_5	------	------	------	------
λ_6	.008	.001	.012	.012
λ_7	.008	.001	.012	.012
λ_8	.008	.001	.012	.012
δ_1	.004	.001	.009	.009
δ_2	.003	.001	.007	.007
δ_3	.005	.001	.006	.006
δ_4	.003	.001	.009	.009
δ_5	.003	.001	.013	.013
δ_6	.003	.001	.011	.011
δ_7	.003	.001	.011	.011
δ_8	.005	.001	.014	.014
ϕ_{12}	.005	.001	.010	.010
ϕ_{11}	.008	.002	.016	.016
ϕ_{22}	.008	.002	.017	.017

Note: *Weighted maximum likelihood using frequency weights in Mplus; **Weighted maximum likelihood using design weights in Lisrel

The second set of analyses was conducted to evaluate the effects of sample size on the estimation techniques because large sample sizes often minimize differences between estimation techniques. Ten randomly selected sub-samples of 250 participants were taken from the full sample for these analyses. The various weighting strategies were recalculated in each of these sub-samples. Once the weighting strategies and the estimation techniques were

applied to each of the 10 sub-samples, the results were averaged across the sub-samples. The results are reported in a similar manner to the first set in Tables 5, 6, and 7.

Table 4
Effective N, Chi-square, and RMSEA for the full sample (N=25,947)

Statistic	UML	WML(M)*	WML(L)**	WRML	WEML	PML
Effective N	25,947	770,512	770,512	26,008	11,971	25,947
Chi-square	4,003.9	114,179.1	1,291.5	1,291.5	1,291.5	1,361.1
RMSEA	.090	.088	.009	.051	.075	.052

Note: *Weighted maximum likelihood using frequency weights in Mplus; **Weighted maximum likelihood using design weights in Lisrel

As in the full sample, the parameter estimates were found to be very similar across the methods. The standard errors for WML(M) estimates are obviously much too small due to the assumed sample size based on the frequency weights. Interestingly, UML standard errors are considerably smaller than PML standard errors. Theoretically, the PML standard errors should be more accurate than the UML standard errors. However, further research is needed to verify this conjecture.

The Chi-square and RMSEA values (Table 7) also show a similar pattern of relative sizes to the results from the full sample analyses. However, the Chi-square value for the WML(M) analysis is much larger than in the full sample and the Chi-square values for the other methods are substantially smaller than in the full sample. The RMSEA values from the WML methods vary greatly and the UML and WML(M) values are larger than the PML value. Although the RMSEA was developed to be relatively unaffected by sample size, the conclusion that researchers might draw from Table 7 differ from those based on Table 4: Whereas PML and WRML appear satisfactory in Table 4, the RMSEA for the PML is 23% larger in Table 7. The WRML RMSEA results are similar in both analyses. Thus, a researcher using the PML technique might conclude that the 2-factor CFA model did not adequately fit in the smaller sample. Only the WML(L) and WRML RMSEAs would suggest that the CFA model fits well.

DISCUSSION

Our goal was to evaluate the effects of UML, WML, and PML estimation techniques on the CFA parameter estimates from large and small stratified samples with informative sampling weights. While the estimation techniques are not strictly comparable due to variations in their Mplus and Lisrel implementations and the assumptions of the techniques, we are able to examine the outcomes of using each technique. The different estimation techniques did not result in substantially different estimates of means, variances, factor loadings, and factor correlation in this sample of 25,947 participants with a range from 0.96 to 183.7 in the raw sample weights. Differences in the parameter estimates were larger in the sub-samples of 250, but the differences were still not large. In general, as sample sizes decreases, the differences in the parameter estimates should on average increase.

The three estimation techniques did differ substantially in their estimates of standard errors of the parameter estimates. In the full sample, the UML and PML techniques resulted in similar standard error estimates; however, in the smaller stratified sample the UML method appears to estimate standard errors that are too small compared with the theoretically preferable PML method. This finding suggests caution when testing the significance of parameter estimates

from small stratified samples when using UML estimation. The WML(M) technique estimates standard errors based on the assumption that the sampling weights are actually frequency weights. In order to enter the sample weights as frequency weights in Mplus, the sample weights had to be integer values, so the frequency weights that resulted were only approximations to the sample weights which had a range of 0.01 to 183.7. The frequency weights yielded an N of 770,641 instead of the correct N of 25,947 in the full sample and an average N of 758,538 instead of the correct N of 250 in the reduced samples. The resulting standard errors of WML(M) were dramatically smaller than they should be. Consequently, Type I error rates of significance tests of parameter estimates would be grossly inflated.

Table 5
Average parameter estimates for 10 reduced (N=250) samples using UML, WML, and PML estimators

	Unstandardized			Standardized		
Statistic	**UML**	**WML****	**PML**	**UML**	**WML****	**PML**
μ_1	3.566	3.635	3.635			
μ_2	3.637	3.692	3.692			
μ_3	3.500	3.580	3.580			
μ_4	3.236	3.279	3.279			
μ_5	2.984	3.010	3.010			
μ_6	2.657	2.646	2.646			
μ_7	2.594	2.599	2.599			
μ_8	2.713	2.733	2.733			
λ_1	1.000*	1.000*	1.000*	.896	.883	.883
λ_2	1.006	.995	.995	.901	.877	.877
λ_3	1.155	1.135	1.135	1.035	.996	.996
λ_4	1.112	1.093	1.093	.996	.958	.958
λ_5	1.000*	1.000*	1.000*	.895	.908	.908
λ_6	1.189	1.166	1.166	1.064	1.057	1.057
λ_7	1.244	1.224	1.224	1.113	1.111	1.111
λ_8	.955	.911	.911	.853	.826	.826
δ_1	.438	.361	.361			
δ_2	.309	.308	.308			
δ_3	.218	.223	.223			
δ_4	.422	.440	.440			
δ_5	.699	.661	.661			
δ_6	.464	.492	.492			
δ_7	.333	.327	.327			
δ_8	.717	.638	.638			
ϕ_{12}	.316	.320	.320	.393	.409	.399
ϕ_{11}	.804	.788	.788	1.000	1.000	1.000
ϕ_{22}	.802	.827	.827	1.000	1.000	1.000

Note: *Fixed value used for identification; **WRML and WEML methods provided identical parameter estimates to WML

Table 6
Average standard error estimates for 10 reduced samples (N=250) using UML, WML, and PML estimators

Statistic	UML	WML(M)*	WML(L)**	PML
μ_1	.070	.001	.098	.098
μ_2	.067	.001	.096	.096
μ_3	.072	.001	.100	.100
μ_4	.075	.001	.107	.107
μ_5	.078	.001	.112	.112
μ_6	.080	.002	.117	.117
μ_7	.079	.001	.115	.115
μ_8	.076	.001	.103	.103
λ_1	------	------	------	------
λ_2	.065	.001	.085	.085
λ_3	.069	.001	.096	.096
λ_4	.074	.001	.107	.107
λ_5	------	------	------	------
λ_6	.093	.002	.124	.124
λ_7	.094	.002	.133	.133
λ_8	.089	.001	.122	.122
δ_1	.047	.001	.121	.121
δ_2	.036	.001	.078	.078
δ_3	.034	.001	.060	.060
δ_4	.048	.001	.086	.086
δ_5	.073	.001	.119	.119
δ_6	.061	.001	.110	.110
δ_7	.056	.001	.100	.100
δ_8	.070	.001	.108	.108
ϕ_{12}	.064	.001	.099	.099
ϕ_{11}	.107	.002	.147	.147
ϕ_{22}	.124	.002	.167	.167

Note: *Weighted maximum likelihood using frequency weights in Mplus; **Weighted maximum likelihood using design weights in Lisrel

The WML(L) technique does not fall victim to the problem of treating sample weights as frequency weights; however, depending on the choice of weighting strategy, different RMSEA values were obtained and different interpretations concerning model fit might be made. It appears relative weighting scheme in Lisrel is the closest approximation to the PML results and appears to be the preferred weighting strategy when using Lisrel.

Table 7
Average effective N, Chi-square, and RMSEA and for 10 reduced samples (N=250)

Statistic	UML	WML(M)*	WML(L)**	WRML	WEML	PML
Effective N	250	758,538	758,538	250	127	250
Chi-square	65.3	252,451.74	32.56	32.6	32.6	40.7
RMSEA	.094	.129	.001	.048	.070	.064

Note: *Weighted maximum likelihood using frequency weights in Mplus; **Weighted maximum likelihood using design weights in Lisrel

CONCLUSION

Our goal in this paper was to examine the utility of the various estimation techniques for typical scale evaluation and development from stratified samples with informative sampling weights. There are a few conclusions that appear warranted. First, in large samples it may be possible to simply ignore the sampling weights or use any scaling of the sample weights and estimation technique in CFA analyses without substantially affecting the parameters estimates. Even in small samples, these estimates are fairly robust. This is comforting given that informative sampling weights are not always available when analyzing large scale survey research datasets.

The results are not so comforting when the standard errors of those estimates are considered. The WML(M) technique, which misinterprets the sample weights as frequency weights, leads to severely underestimated standard errors for parameter estimates when the degree of stratification is moderate. This result depends upon the sampling design rather than the sample size. The UML standard errors, which are estimated when sampling weights are ignored, appear to be downwardly biased (Kaplan & Ferguson, 1999). This suggests caution when making inferences from these estimates. The WML techniques available in Lisrel replicated the standard errors from the PML technique and therefore seem preferable; however, the RMSEA values for these techniques were not equivalent.

The RMSEA values from the WRML weighting scheme most closely approximated the PML results in the full sample, and the RMSEA values from the WEML weighting scheme most closely approximated the PML results in the full reduced sample. So, given the equivalence on the parameter estimates and standard errors for the other weighting schemes, either the relative or the effective weights may be appropriate when using Lisrel for estimating population parameters from stratified samples; however, the RMSEA values for the WRML were smaller than the PML values, which would lead to less conservation inferences about model fit. When evaluation of overall model fit is of great importance, it may be worth the time and expense to use the PML technique in Mplus Version 3.0.

In sum, the advantage of using the more complicated methods of CFA estimation will vary depending upon the sample size and the importance of the accuracy of the parameter estimates beyond the first decimal point. Moreover, it is almost always important to interpret parameter estimates in light of their standard errors and the extent to which the overall fit of the model is satisfactory. Consequently, our analyses suggest that the WRML, WEML, and PML methods provide the most trustworthy results for stratified random samples such as the 2004 Workplace Gender Relations survey.

REFERENCES

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12,* 411-434.

Brown, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136-162). Newbury Park: Sage Publications

Flores-Cervantes, I., Jones, M. E., & Wilson, M. J. (2004). Weighting for the 2004 Workplace and Gender Relations Survey of Reserve Component Members. In R. A. Riemer (Ed.), *2004 Workplace and Gender Relations Survey of Reserve Component Members: Statistical Methodology Report* (Report No. 2004-019). Arlington, VA: DMDC

Gade, P. A. Tiggle, R. B; & Schumm, W. R. (2003). The measurement and consequences of military organizational commitment in soldiers and spouses. *Military Psychology, 15,* 191-207.

Kaplan, D. & Ferguson, A. J. (1999). On the utilization of sample weights in latent variable models. *Structural Equation Modeling, 6,* 305-321.

Kroeger, K. R. (2004). Sample design for the 2004 Workplace and Gender Relations Survey of reserve component members. In R. A. Riemer (Ed.), *2004 Workplace and Gender Relations Survey of Reserve Component Members: Statistical Methodology Report* (Report No. 2004-019). Arlington, VA: DMDC.

Jöreskog, K.G. & Sörbom, D. (2004). ***LISREL 8.7*** *for Windows [Computer Software]*. Lincolnwood, IL: Scientific Software International, Inc.

Meyer, J.P., & Allen, N.J. 1990. The measurement and antecedents of affective, continuance, and normative commitment to the organization. *Journal of Occupational Psychology,* 63, 1-18.

Muthén, B. O. (1998-2004). *Mplus technical appendices.*. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K. (2005, March 2). Mplus discussion. Message posted to http://www.statmodel.com/discussion/messages/9/579

Muthén, L. K. and Muthén, B. O. (1998-2004). *Mplus user's guide*. Third Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K. and Muthén, B. O. (2004). *Mplus 3.0 [Computer Software]*. Los Angeles, CA: Muthén & Muthén.

Ormerod, A. J., Lawson, A. K., Lytell, M. C., Wright, C. V., Sims, C. S., Brummel, B.J., Drasgow, F., Lee, W. C., & Fitzgerald, L. F. (2005). *2004 Workplace and Gender Relations Survey of Reserve Component Members: Scales and measures report* (Report No. 2004-022). Arlington, VA: DMDC.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 60,* 23-40.

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association*, 87, 383-396.

Riemer, R. A. (Ed.). (2004). *2004 Workplace and Gender Relations Survey of Reserve Component Members: Statistical Methodology Report* (Report No. 2004-019). Arlington, VA: DMDC.

Skinner, C. J. (1989). Domain means, regression, and multivariate analysis. In C. J. Skinner, D. Holt, and T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59-87). New York: Wiley.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). General Introduction. In C. J. Skinner, D. Holt, and T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 5-20). New York: Wiley.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9,* 475-502.

Applied Multivariate Research, Volume 13, No. 2, 2010, 129-144

AN ITEM RESPONSE THEORY ANALYSIS OF THE SELF-MONITORING SCALE

Edward Burkley[1]
Oklahoma State University

ABSTRACT

The Self-Monitoring Scale (SMS) was investigated utilizing item response theory (IRT). First, IRT models that constrained each of the subscale items to have equal discrimination were fitted to the three subscales of the SMS (Acting, Extraversion, and Other-Directedness). These models were then contrasted with separate models that allowed the discriminations to be estimated freely. For all three subscales, model comparison tests of significance indicated that the unconstrained models were a better fit. Thus, the items of each subscale are differentially related to their respective underlying construct. Implications and recommendations are offered for future psychometric development and implementation of the SMS.

Keywords: Self-monitoring, Item Response Theory

INTRODUCTION

The ability of expressive control over one's behavior from situation to situation has important implications for the link between the person and the situation. Variables that moderate this relationship between personality and social behavior are among the most useful constructs in personality and social psychology (Snyder & Ickes, 1985). One such variable that has received a great deal of attention is that of self-monitoring (Snyder, 1974). Since its introduction over 30 years ago, the construct of self-monitoring has spawned a cornucopia of research and has been claimed as "one of the most popular measures of personality to be introduced in recent years" (Briggs & Cheek, 1988, p. 663). Given the importance of a variable like self-monitoring and its many uses in varying research areas, numerous psychometric investigations have been conducted to evaluate the most popular measure of self-monitoring: The Self-Monitoring Scale (SMS; Snyder, 1974). Yet, no detailed item analysis has been performed to improve the psychometric development and implementation

[1] Correspondence concerning this article should be addressed to Edward Burkley, Department of Psychology, Oklahoma State University, 116 N. Murray, Stillwater, OK 74078; email: ed.burkley@okstate.edu.

of the SMS. Thus, one potential avenue of investigation is to utilize the benefits of item response theory techniques (IRT; Embertson & Reise, 2000; Thissen & Wainer, 2001; van der Linden & Hambleton, 1997).

The purpose of the present article is to utilize IRT and offer a more refined item analysis of the SMS than is feasible with other more traditional psychometric techniques. Although there is overlap between classic testing approaches and IRT, IRT does offer some advantages for those interested in personality assessment (Reise & Henson, 2003). For example, by using IRT to analyze the Rosenberg Self-Esteem scale, Gray-Little, Williams, and Hancock (1997) were able to determine that this measure is quite poor at differentiating among high self-esteem individuals. That is, this measure is useful for distinguishing between people low in self-esteem from those that are average, but is not adequate for distinguishing between people who are high in self-esteem. For any researcher interested in using this self-esteem scale, this information may be critical. In another example, Reise and Henson (2000) demonstrated that one item of the Revised NEO-PI Anxiety subscale was nearly four times as informative as the rest of the items. Both of these examples demonstrate situations in which an IRT analysis offered novel evidence above and beyond what could be obtained using a more traditional psychometric approach.

In this article, I will briefly outline the history of self-monitoring assessment. I will then provide an overview of item response theory and explain how I applied this technique to the primary tool for assessing self-monitoring (i.e., the self-monitoring scale and its three subscales). After presenting the results, I will discuss implications and recommendations regarding the future psychometric development and implementation of the self-monitoring scale.

Evolution of the Self-Monitoring Scale

Self-monitoring theory asserts that people vary in the extent to which they regulate their expressive behavior (Snyder, 1974). People who are *high self-monitors* exert a great deal of control over their behaviors out of a concern for situational appropriateness. The behavior of high self-monitors is largely determined by the social constraints of the situation. Consequently, they are often referred to as "social chameleons". People who are *low self-monitors* do not engage in such self-presentation. Instead, their behavior represents their own inner attitudes and emotions. Accordingly, the behavior of low self-monitors is less influenced by the situation and is more reflective of the individual's disposition. This distinction between those receptive to the social environment versus those reliant on inner qualities gives self-monitoring the favorable quality of being able to serve as an important moderator of social behavior.

To measure the construct of self-monitoring, Snyder developed the Self-Monitoring Scale (SMS; Snyder, 1974). The scale is comprised of 25 items in a true-false response format. Since the scale's inception, numerous research investigations have utilized the scale and its importance cannot be overstated. The majority of these investigations have focused on the factor structure of the SMS. Investigators have identified factor solutions ranging from one to four, with three factors being the most widely accepted solution (Briggs, Cheek, & Buss, 1980; Gangestad & Snyder, 1985). These three factors were labeled *Acting*, *Extraversion*, and *Other-Directedness*. However, Snyder and Gangestad (1986) argued that this rotated three-factor solution is not as informative in identifying a single latent construct. To improve the psychometric properties of the scale, Snyder and Gangestad (1986) offered a revised SMS comprised of 18 of the SMS items that had the largest factor loadings. Later investigations

contrasted the 18-item SMS with the original SMS (Briggs & Cheek, 1988; Hoyle & Lennox, 1991; John, Cheek, & Klohnen, 1996), revealing that this revised SMS has its own set of issues. Generally, it has been demonstrated that the deletion of the items from the original SMS increased the scale's reliability and purity of the factors, yet it shifted the scale toward Extraversion and Acting at the expense of weakening the Other-Directedness factor (Hoyle & Lennox, 1991; John et al., 1996). Based on this evidence it has been recommended that researchers restrict investigations of self-monitoring to the use of the original 25-item scale with its three subscales: Acting, Extraversion, and Other-Directedness (e.g., John et al., 1996).

What is needed in the investigation of the psychometric properties of the SMS is a method 1) that is well suited for the use of dichotomous data (true/false), the typical format used when administering the SMS and 2) that does not make unsound assumptions about item response distributions. Such a resolution is offered by the use of IRT. Not only does IRT fit the above criteria, it also offers a more refined, item-by-item analysis of the scale. First, item response models were originally designed for use on dichotomous data, although more sophisticated models are now available that handle ordered/unordered categorical data (van der Linden & Hambleton, 1997). Second, IRT offers a thorough item-level analysis, providing more information than other traditional techniques. For example, IRT can indicate which items of the SMS do a better job of differentiating between individuals who are low and high self-monitors. Thus, IRT offers a unique contribution to the understanding of the SMS by providing valuable information about the individual items and suggesting possible improvements for the overall scale.

Item Response Theory

Terminology and assumptions. Item Response Theory (IRT) is a collection of statistical methods that assesses the extent to which each item measures the underlying latent construct. This latent construct is generally referred to as theta (θ) and is typically distributed as a z-score that ranges from –3 to +3.

There are two parameters that are central to any IRT model - the threshold (*b*) and discrimination (*a*) parameters. The *threshold parameter (b)* locates the level of the underlying construct where there is a .50 probability of endorsing the item. For personality scales, this parameter indicates the level of the personality trait necessary to endorse the item. Thus, SMS items with large positive *b* values are only endorsed by individuals with very high levels of self-monitoring. These items would distinguish between moderate and high self-monitors. Conversely, SMS items with large negative *b* values are endorsed by individuals with very low levels of self-monitoring. These items would distinguish between moderate and low self-monitors.

The item *discrimination parameter (a)* quantifies the association between the item and the latent construct. It represents the item's ability to discriminate among people with different levels of the underlying trait. The *a* parameter typically ranges from 0 to 3. SMS items with large *a* values do an excellent job of discriminating between various levels of self-monitoring, whereas SMS items with small *a* values indicate that individuals' responses to these items do not relate to the self-monitoring trait. Thus, the *b* parameter indicates *where* on the continuum the item is discriminating; whereas, the *a* parameter indicates *how well* the item is discriminating.

An additional piece of information unique to IRT is a scale's *marginal reliability*. Although reliability is a classic test theory concept, marginal reliability is an analogous index

that is often estimated when using IRT. The values for marginal reliability range from 0 to 1 and their interpretation is analogous to Cronbach's alpha (Green, Bock, Humphreys, Linn & Reckase, 1984). This statistic is useful because it provides a single numeric value that summarizes the scale's overall precision (Thissen & Wainer, 2001).

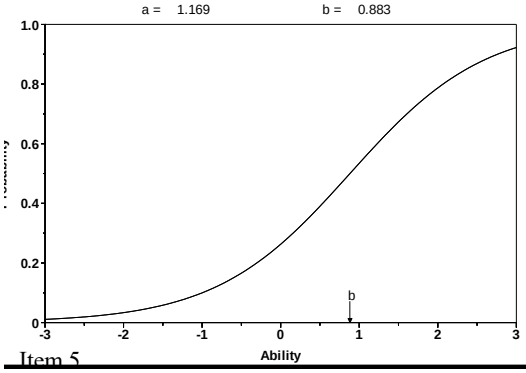
IRT is based on two major assumptions: Unidimensionality and local independence. The assumption of unidimensionality states that all items being analyzed reflect a single continuous latent construct (θ). When a scale consists of multiple subscales, as does the SMS, each subscale is treated as a unidimensional construct. The assumption of *local independence* states the items are independent of each other and that the only connection that does exist between the items is in regards to their relationship with the latent construct. For example, items that are grouped in a testlet would not reflect local independence because there is an additional connection that exits between the items (i.e., grouped items directed toward a single topic). Before conducting an IRT analysis, exploratory or confirmatory factor analysis is typically used to evaluate the unidimensionality and local independence of a particular scale or subscale. In practice, a scale rarely meets these assumptions perfectly; however, IRT is known to be relatively robust to moderate violations of these assumptions (e.g., Drasgow & Parsons, 1983; Hulin, Drasgow, & Parsons, 1983).

Another important note is that IRT relies heavily on graphic representations of item characteristics. The cornerstone of IRT is the *Item Characteristic Curve* (ICC) or trace line, which represents the probability of an item response as a function of the underlying construct. The ICC is central to IRT because it can be used to evaluate the quality of each item. In this graph, the underlying trait comprises the *x*-axis and the probability of endorsement comprises the *y*-axis (see Figure 1 for an example). For each item, the slope of the line represents the item discrimination parameter (a), with steeper (i.e., larger) slopes indicating greater discrimination ability.

Finally, an important concept in IRT is that of the information function. The *item information function* is an index of how much psychometric information the item provides at each level of the latent construct. In other words, this information function represents how well each item differentiates between people at different levels of the latent construct. More importantly, item information functions are useful because they can be combined to form an overall test information function. The *test information function* is useful in that it indicates how well the entire scale differentiates between people at different levels of the latent construct. For example, a particular self-esteem scale may be more precise at assessing people high in self-esteem; whereas another self-esteem scale may be better equipped at assessing people low in self-esteem. The graphical representation of information functions places the underlying trait on the *x*-axis and the amount of information on the *y*-axis (see Figure 4 for example of test information functions). The peak of the information curve indicates where on the theta continuum the scale has the greatest amount of precision, or information. These graphs offer an additional piece of data that is also quite useful, as information increases (solid line), measurement error can be seen to decrease (dotted line), thus offering an assessment of precision and its inverse, error.

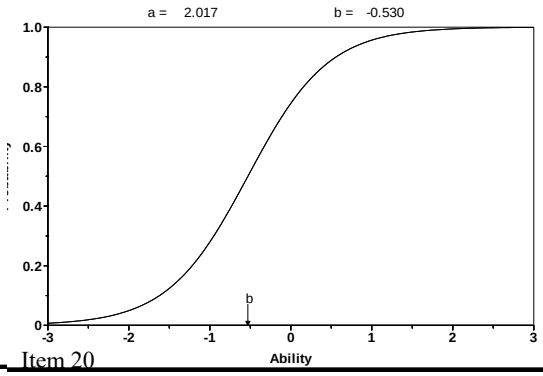
Figure 1

Item characteristic curves for the Acting subscale (2PL model). The x-axis represents the latent construct of acting (θ). The y-axis represents the probability of a true response [T(x = 1|θ)]. Note. * indicates reverse scored items.



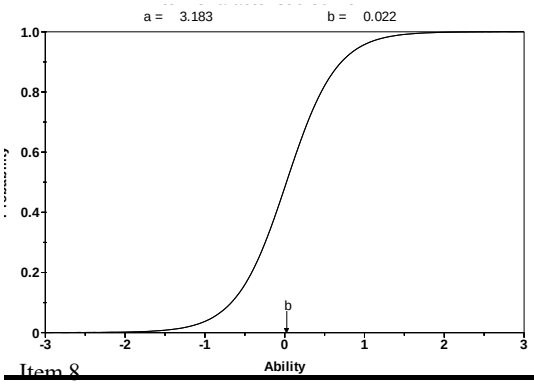
Item 5

I can make impromptu speeches even on topics about which I have almost no information.



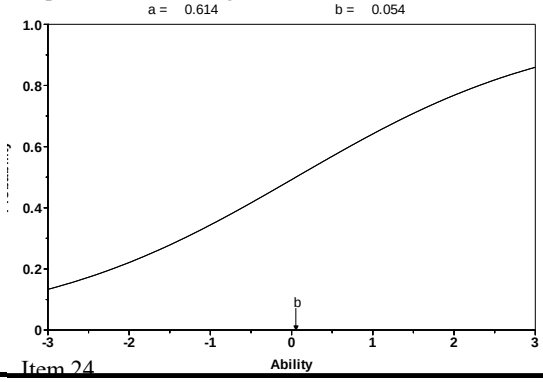
Item 20

I have never been good at games like charades or improvisational acting.*



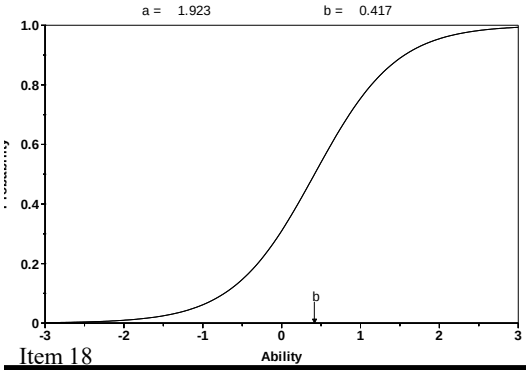
Item 8

I would probably make a good actor.



Item 24

I can look anyone in the eye and tell a lie with a straight face (if for a right end).



Item 18

I have considered being an entertainer.

IRT models. Although all IRT analyses assume a single underlying construct, a variety of models can be used to define the causal relationship between this construct and the observed item responses. In essence, these models differ in terms of how many parameters are required to model each item response. A review of all IRT models is beyond the scope of this paper, so instead only the models typically used for scales with dichotomous test items will be described (i.e., logistic models).

One of the simplest IRT models is the *1-parameter logistic model* (1PL). In this model, only the threshold parameter (*b*) is required to model the item process. This model is often referred to as a restricted model because it restricts the *a* parameters to be equal across all items. By imposing this restriction, this model makes the assumption that all the items are equally related to their underlying construct and therefore discriminate equally.

For item *j*, the 1PL model for the trace line is defined as:

$$T_j(x_j = 1|\theta) = \{1 + \exp[-a(\theta - b_j)]\}^{-1}.$$

In this equation, $T_j(x_j = 1|\theta)$ traces the probability of a "true" response ($x_j = 1$) as a function of θ, b_j represents the threshold parameter for item *j* and indicates at what level of θ an individual has a 50% chance of giving a "true"(or high trait) response to item *j*. Finally, *a* is the item discrimination parameter (slope) and represents the rate of change in the proportion of "true" responses as a function of θ. As stated above, this equation allows the *b* parameters to be assessed but restricts the *a* parameters to be equal across all items.

Although the 1PL model is informative, in practice, scale items do not typically discriminate equally. To address this issue, a *2-parameter logistic model* (2PL) is often employed. The 2PL model is almost identical to the 1PL except that the discrimination parameter (*a*) is allowed to vary across items (i.e., an unrestricted model). This model therefore allows the items to vary in their relation to the underlying construct. The equation for the 2PL model is identical to the one expressed above, with the addition of a *j* subscript to the *a* parameter thereby allowing the *a* parameter to vary across items.

When utilizing IRT models, it is common practice to compare the results of the 1PL and 2PL models to determine if each item is equally related to the latent construct. By examining the difference in fit between the 1PL and the 2PL models, one is able to empirically test the assumption that all scale items relate equally to the underlying trait. The assessment of this assumption is important to test because it is what allows researchers to sum across all items to create a composite score (this is standard procedure for the SMS). If this assumption is violated and items do differ in their relation to the underlying trait, then summed scores are an erroneous practice for that particular scale.

Analyses of the SMS Subscales. Past investigations of the SMS using factor analysis have suggested a multifactorial structure (e.g., Briggs et al., 1980; Hoyle & Lennox, 1991; John et al., 1996). The most often agreed upon solution posits three factors, generally labeled as Acting, Extraversion, and Other-Directedness (Briggs et al., 1980). To examine the psychometric properties of each subscale, separate IRT analyses were conducted on each of the three SMS subscales.

METHOD

Data

The data were collected from introductory psychology students from a large southeastern university. The sample consisted of 581 students who completed the original 25-item SMS in a true-false format. The scale was scored so that true responses received a 1 and false responses received a 0. The items were then recoded so that a 1 represented a "high self-monitoring response" and 0 represented a "low self-monitoring response."

Based on previous SMS research (Briggs et al., 1980), the subscales were defined as follows: The Acting subscale consisted of items 5, 8, 18, 20, and 24 (see Figures 1, 2, and 3 for items), the Extraversion subscale consisted of items 12, 14, 20, 21, 22, and 23, the Other-Directedness subscale consisted of items 2, 3, 6, 7, 13, 15, 16, 17, 19, 23, and 25. Items 1 and 4 were not included in any of the subscales (Briggs et al., 1980).

Models

The three subscales of the SMS (Acting, Extraversion, and Other-Directedness) were analyzed using binary IRT models in the Multilog computer program (Thissen, 1991; Thissen, Chen, & Bock, 2003). First a 1PL model was fit to the data, restricting all items to be equally related to their underlying construct. Next, a 2PL model was fit to the data, allowing the items to vary in their relation to the underlying construct.

RESULTS

Test of Assumptions

A confirmatory factor analysis with three uncorrelated factors representing each subscale was conducted to assess unidimensionality and local independence. Because the data were dichotomous, a robust weighted least squares (WLSMV) estimation was conducted in Mplus (Muthén & Muthén, 2000). The *RMSEA* demonstrated good fit (*RMSEA* = .06) but the *TLI* fell just short of the desirable cutoff (*TLI* = .85). Although these results are not conclusive, there is a great deal of research demonstrating evidence for the prominent 3-factor solution of the SMS and there is also evidence that IRT is relatively robust to moderate violations of its assumptions (e.g., Drasgow & Parsons, 1983; Hulin et al., 1983); thus, IRT analyses were pursued with an assessment of each of the SMS subscales.

Acting Subscale

Descriptive statistics. The average summed score on the Acting subscale was 2.32 (*SD* = 1.58), with scores ranging from 0 to 5. The means for the acting items were .31, .49, .38, .66, and .49, respectively. The coefficient alpha was .66 and the mean item-to-total correlation (r_{φ}) was .67. The interitem correlations ranged from $r_{\varphi\ 20,24} = .10$ to $r_{\varphi\ 8,20} = .46$ and the item-total correlations ranged from $r_{\varphi\ T,24} = .55$ to $r_{\varphi\ T,8} = .73$.

IRT analysis. First, the 1PL model was applied to the data. This restricted model estimated 6 parameters. The *-2log*(likelihood) computed for this model was –3868.10 and the marginal reliability was .63. Next, the 2PL model was applied to the data. This unrestricted model estimated 10 parameters. The *-2log*(likelihood) was –3923.30 and the marginal reliability was

.66. The relative fit of the unrestricted and restricted models was then assessed for the Acting subscale,

$G_{diff}^2 =$ -3868.10 - (-3923.30) = 55.20 ($df = 4$, $p < .001$).

The results indicated that the 2PL model fit the data significantly better than the 1PL model. Thus, one can conclude that the discrimination parameters are not equal across items and that these subscale items differ in their relation to acting.

Table 1
Parameter Estimates for the Self-Monitoring Subscales (2PL model)

Subscale			
Acting		*a* (*SE*)	*b* (*SE*)
	Item 5	1.17 (0.16)	0.88 (0.14)
	Item 8	3.18 (0.36)	0.02 (0.05)
	Item 18	1.92 (0.22)	0.42 (0.08)
	Item 20	2.02 (0.22)	-0.53 (0.07)
	Item 24	0.61 (0.12)	0.05 (0.18)
Extraversion			
	Item 12	1.86 (0.20)	-0.06 (0.07)
	Item 14	1.36 (0.18)	-1.33 (0.16)
	Item 20	0.96 (0.14)	-0.80 (0.15)
	Item 21	0.59 (0.13)	-1.58 (0.36)
	Item 22	1.88 (0.20)	-0.04 (0.07)
	Item 23	1.73 (0.20)	-0.64 (0.09)
Other-Directedness			
	Item 2	0.87 (0.16)	1.78 (0.30)
	Item 3	0.57 (0.14)	-2.29 (0.73)
	Item 6	0.90 (0.14)	0.56 (0.35)
	Item 7	0.47 (0.15)	-3.82 (1.59)
	Item 13	1.30 (0.16)	-0.48 (0.10)
	Item 15	0.93 (0.14)	-0.15 (0.13)
	Item 16	1.29 (0.20)	-0.76 (0.12)
	Item 17	0.41 (0.11)	1.30 (0.42)
	Item 19	1.48 (0.23)	1.26 (0.14)
	Item 23	0.07 (0.12)	-10.69 (* *)
	Item 25	0.75 (0.13)	-0.71 (0.19)

* * Unable to obtain estimate of standard error, value too large.

The threshold and discrimination parameters for the 2PL model are displayed in Table 1 and the ICCs for each item are shown in Figure 1. The threshold parameters ranged from b_{20} = -.53 to b_5 = .88. All of the items' threshold values were close to zero, where zero on the theta continuum represents the unobserved population mean. This suggests that this subscale allows fine distinctions only among individuals with moderate levels of the acting construct. These items are only providing information about moderate levels of acting and do not supply much information regarding the lower and upper ends of the acting continuum. This indicates that the use of this subscale does not allow fine distinctions among individuals with low or high levels of acting. This pattern can also be seen graphically in the test information function (see Figure 4), with the greatest amount of measurement precision, and the least amount of error, occurring in the center of the continuum.

The item discriminations ranged from a poor a_{24} = .61 to a very strong a_8 = 3.18. This variability can also be seen in the ICCs for these items. Items 8, 18, and 20 show strong slopes, with item 8 displaying the steepest slope. Thus, items 8, 18, and 20 show the strongest relation to acting. Conversely, item 24 shows a weaker relationship to acting, as indicated by its flatter slope.

Extraversion Subscale

Descriptive statistics. The average summed score on the Extraversion subscale was 3.86 (*SD* = 1.68), with scores ranging from 0 to 6. The means for the extraversion items were .52, .80, .66, .70, .51, and .67, respectively. The coefficient alpha was .64 and the mean item-to-total correlation (r_φ) was .61. The interitem correlations ranged from $r_{14,21}$ = .07 to $r_{12,22}$ = .40 and the item-total correlations ranged from $r_{T,21}$ = .48 to $r_{T,(12,\,22)}$ = .66.

IRT analysis. First, the 1PL model was applied to the data. This restricted model estimated 7 parameters. The *-2log*(likelihood) computed for this model was –3310.80 and the marginal reliability was .61. Next, the 2PL model was applied to the data. This unrestricted model estimated 12 parameters. The *-2log*(likelihood) computed for this model was –3344.60 and the marginal reliability was .63. The relative fit of the unrestricted and restricted models was then assessed for the Extraversion subscale,

G_{diff}^2 = -3310.80 - (-3344.60) = 33.80 (df = 5, p < .001).

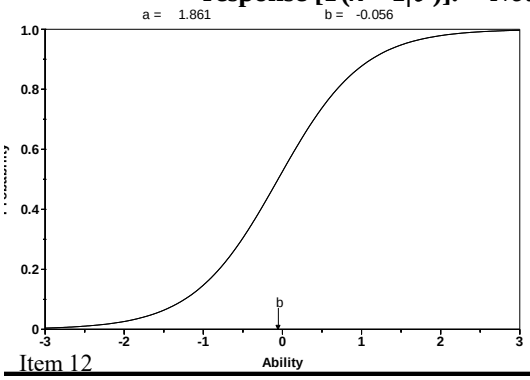
The results indicated that the 2PL model fit the data significantly better than the 1PL model. Thus, one can conclude that the discrimination parameters are not equal across items and that these subscale items differ in their relation to extraversion.

The threshold and discrimination parameters for the 2PL model are displayed in Table 1 and the ICCs for each item are shown in Figure 2. The threshold parameters ranged from b_{21} = -1.58 to b_{22} = -.04. All of the items' threshold values were negative, indicating that this subscale allows fine distinctions only among individuals with low levels of extraversion and that none of the items are adequately identifying those individuals high in extraversion. This pattern is reiterated graphically in the test information function (see Figure 4), with the greatest amount of measurement precision occurring in the center and lower end of the continuum.

The item discriminations ranged from a_{21} = .59 to a_{22} = 1.88. As shown in the ICCs, items 12, 22, and 23 show the strongest slopes, and thus are most related to the construct of extraversion. Item 21 has the weakest relation to extraversion, as is evident by its flatter slope.

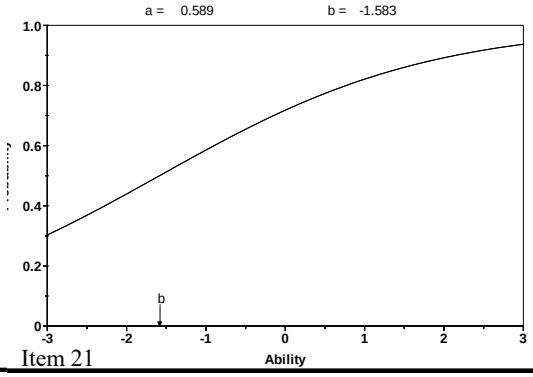
Figure 2

Item characteristic curves for the Extraversion subscale (2PL model). The x-axis represents the latent construct of extraversion (θ). The y-axis represents the probability of a true response [$T(x = 1|\theta)$]. *Note.* * indicates reverse scored items.



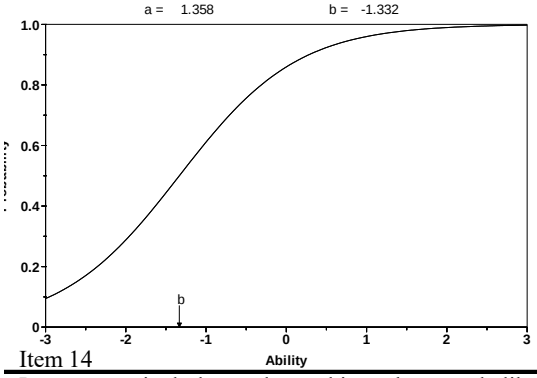
Item 12

In a group of people I am rarely the center of attention.*



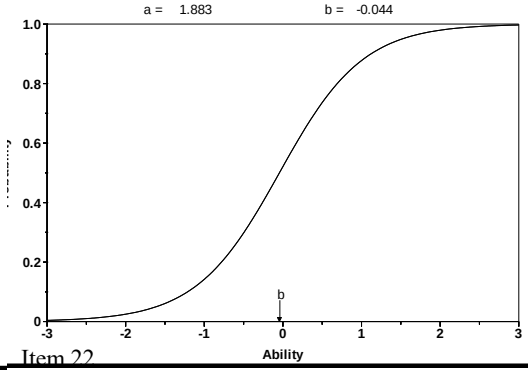
Item 21

I have trouble changing behavior to suit different people and different situations.*



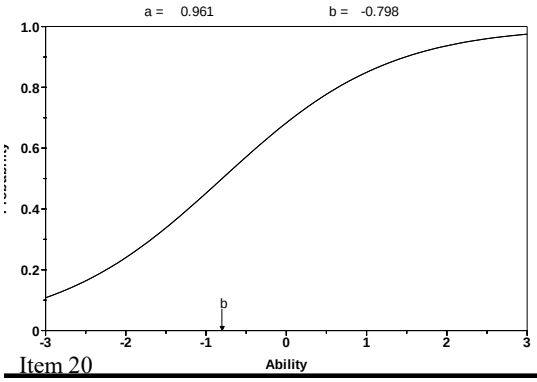
Item 14

I am not particularly good at making other people like me.*



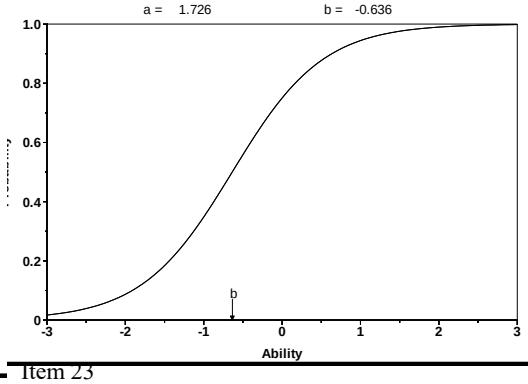
Item 22

At a party I let others keep the jokes and stories going.*



Item 20

I have never been good at games like charades or improvisational acting.*



Item 23

I feel a bit awkward in company and do not show up quite as well as I should.*

Other-Directedness Subscale

Descriptive statistics. The average summed score on the Other-Directedness subscale was 5.92 (*SD* = 2.03), with scores ranging from 0 to 11. The means for the other-directedness items were .21, .77, .39, .85, .62, .53, .68, .38, .20, .67, and .62, respectively. The coefficient alpha was .50 and the mean item-to-total correlation (r_{φ}) was .45. The interitem correlations ranged from $r_{7,23} = .001$ to $r_{13,16} = .28$ and the item-total correlations ranged from $r_{T,23} = .10$ to $r_{T,13} = .52$.

IRT analysis. First, the 1PL model was applied to the data. This restricted model estimated 12 parameters. The *-2log*(likelihood) was 117.40 and the marginal reliability was .50. Next, the 2PL model was applied to the data. This unrestricted model estimated 22 parameters. The *-2log*(likelihood) computed for this model was 7.6 and the marginal reliability was .60. The relative fit of the unrestricted and restricted models was then assessed for the Other-Directedness subscale,

$G_{diff}^2 = 117.40 - 7.6 = 109.80$ ($df = 10$, $p < .001$).

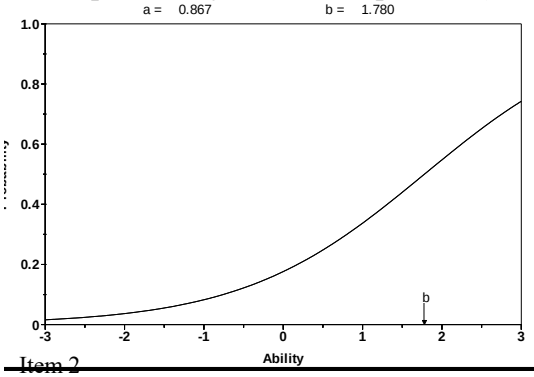
Once again, the results indicated that the 2PL model fit the data significantly better than the 1PL model. Thus, the discrimination parameters are not equal for all items, indicating that these subscale items differ in their relation to other-directedness.

The threshold and discrimination parameters for the 2PL model are displayed in Table 1 and the ICCs for each item are shown in Figure 3. The threshold parameters ranged from $b_2 = 1.78$ to $b_{23} = -10.69$. The majority of the items had negative threshold values, indicating that this subscale allows for fine distinctions only among individuals with low levels of other-directedness. Two of the items (7 and 23) showed extremely negative values, indicating that only a very low level of other-directedness would be needed to endorse these items. Items such as these are less informative because most individuals are likely to endorse them and therefore they do not distinguish among different levels of other-directedness.

The item discriminations ranged from a very poor $a_{23} = .07$ to a moderate $a_{19} = 1.48$. This wide variability can also be seen in the ICCs for these items. None of the items show strong slopes. Three of the items (13, 16, and 19) show moderate slopes, with item 19 displaying the steepest slope in the subscale. Thus, item 19 shows the strongest relation to other-directedness, albeit only modestly. Conversely, items 3, 7, 17, and 23 show an extremely weak relationship to other-directedness, as indicated by their very flat slopes. This finding suggests that these items are poor measures of other-directedness and may be related to a different construct. The test information function (see Figure 4) also suggests that this is the case, given that the information curve does not appear to peak and the error curve remains high across the theta continuum.

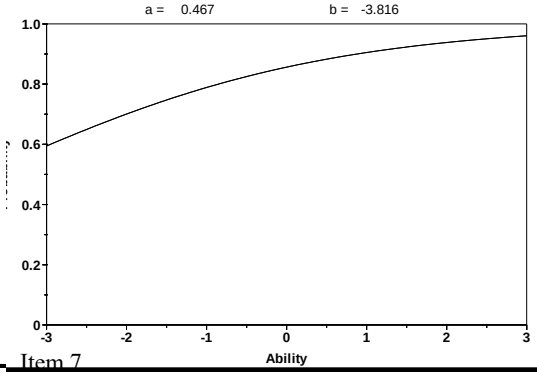
Figure 3

Item characteristic curves for the Other-Directedness subscale (2PL model). The x-axis represents the latent construct of other-directedness (θ). The y-axis represents the probability of a true response [$T(x = 1|\theta)$]. *Note.* * indicates reverse scored items.



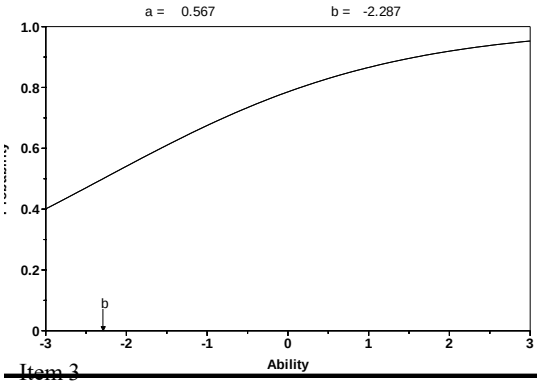
Item 2

My behavior is usually an expression of my true inner feelings, attitudes, and beliefs.*



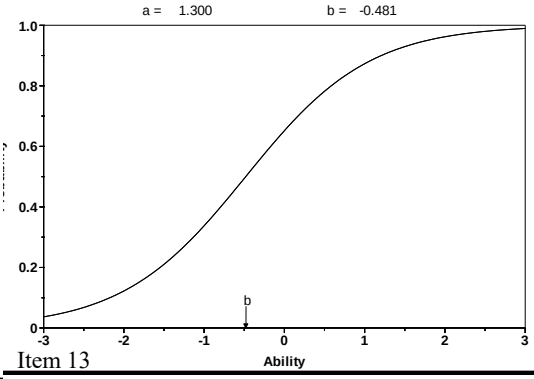
Item 7

When I am uncertain how to act in a social situation, I look to the behavior of others for cues.



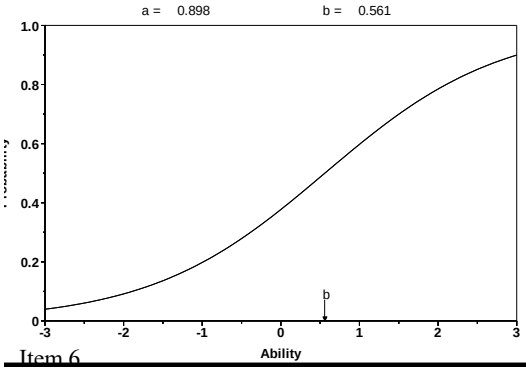
Item 3

At parties and social gatherings, I do not attempt to do or say things that others will like.*



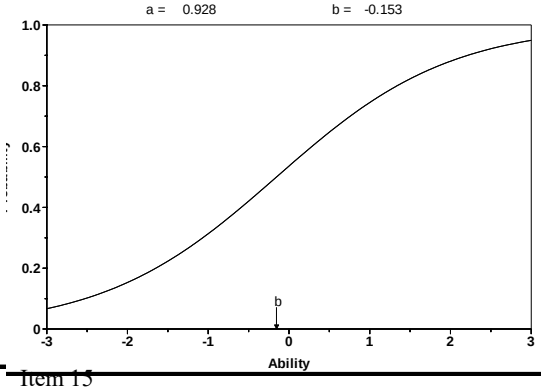
Item 13

In different situations and with different people, I often act like very different persons.



Item 6

I guess I put on a show to impress or entertain people.

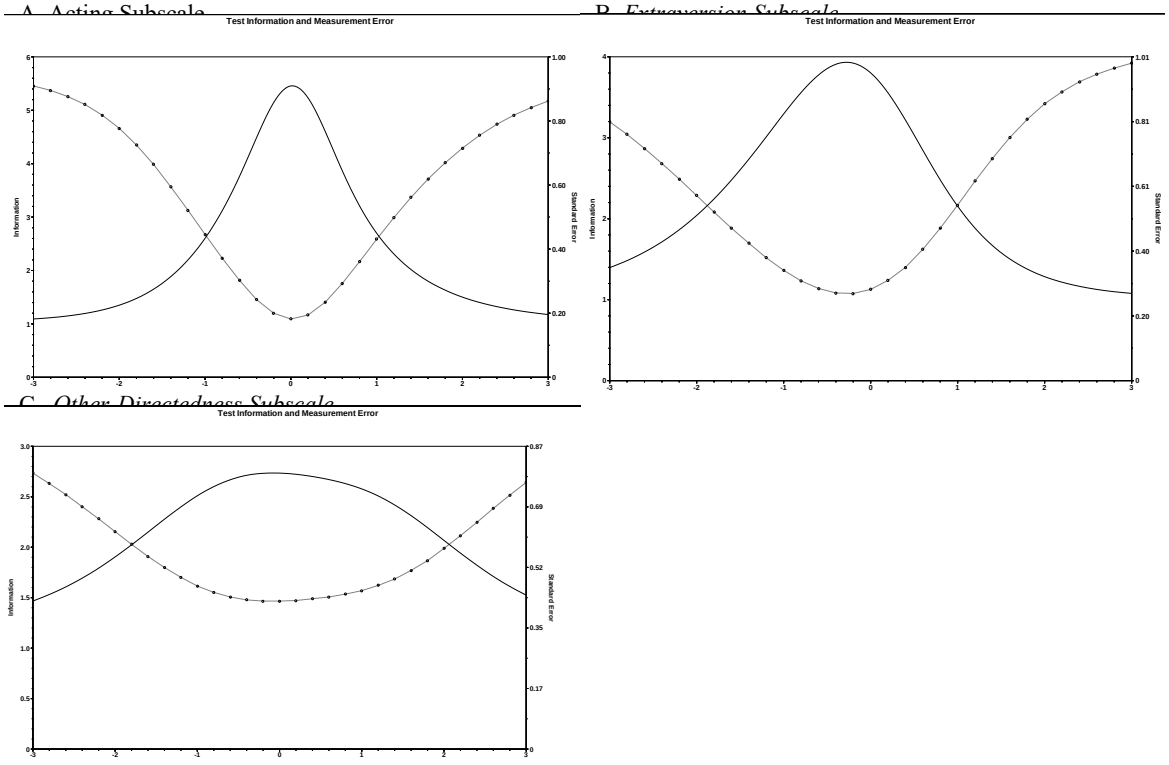


Item 15

Even if I am not enjoying myself, I often pretend to be having a good time.

Figure 4

Test information functions for the Acting, Extraversion and Other-Directedness subscales (2PL models). The x-axis represents the latent construct for each subscale (θ). The left y-axis represents the amount of information provided by each subscale (solid line) and the right y-axis represents the amount of error (dotted line).



DISCUSSION

Summary

This paper represents the first item response theory analysis of the SMS subscales and thus offers an important contribution to the understanding of the Self-Monitoring Scale's psychometric properties. Importantly, the IRT analyses revealed aspects about the scale that mirror previous findings yet it also offered new information, clarification, and suggestions about its use.

Given that some researchers have suggested solely using the SMS subscales (e.g., John et al., 1996), separate IRT analyses were conducted for each subscale. For all three subscales, the IRT analyses showed that the unrestricted model provided a superior fit to the data. Importantly, this indicates that the subscale items are not equal in their relationship to their respective underlying latent construct. Further inspection of the item parameters demonstrated that some subscales displayed greater variability than others, with the Other-Directedness subscale displaying the greatest amount of variability.

The Acting subscale possessed the best psychometric properties. Three of the five items showed very large slopes, indicating that the majority of the items are highly related to the construct of acting. The IRT analysis revealed that items 8, 18, and 20 best captured the acting construct. Further, the thresholds of the items were all close to zero, indicating that these items are differentiating between individuals in the middle of the acting continuum. This property may be desirable when the construct is thought to be normally distributed in the population.

Next, the Extraversion subscale showed reasonable psychometric properties. Three of the six items showed strong slopes, indicating that half of the items are highly related to the construct of extraversion. Items 12, 22, and 23 best captured the extraversion construct. The thresholds of the items were all negative, indicating that the items are primarily identifying individuals low in extraversion. That is, the Extraversion subscale is adequate for identifying people who are low in extraversion. However, this subscale is inadequate for researchers interested in distinguishing among people who are high in extraversion. This information is critical for researchers interested in studying highly extraverted individuals using the SMS.

Finally, the Other-Directedness subscale demonstrated poor psychometric properties. None of the items had strong slopes. Out of the eleven items, only three showed moderate slopes, four of the items had extremely weak slopes, and the rest were in between this range. This suggests that none of these items allow for fine distinctions within the construct of other-directedness. It is important to note that the item with the strongest relation (item 19) was deleted from the original 25-item SMS to create the revised 18-item scale (Snyder & Gangestad, 1986). As previously stated, many argue that this revision shifted the SMS toward Extraversion and Acting and weakened the Other-Directedness construct (Hoyle & Lennox, 1991; John et al., 1996, Briggs & Cheek, 1988). The present analyses reveal that this is most likely because the revised SMS removed the strongest item of the Other-Directedness subscale. Finally, the thresholds of these items were mostly negative, with two items having extremely negative thresholds. Because many of the items do not strongly relate to the construct, their thresholds are more variable and less informative. Upon examination of the few items that did show a moderate relation to other-directedness (i.e., items 13, 16, and 19), one can see that two of the thresholds are negative and one is positive. Thus, together these items are differentiating the upper and lower ends of the continuum.

In sum, the present IRT analysis suggest that the Acting subscale has good psychometric properties, the Extraversion subscale is only useful in identifying Introverts, and the Other-Directedness subscale is a poor measure of its construct. Researchers interested in using the SMS should be aware of these limitations so that they can decide if this measure is appropriate for their particular research needs.

Implications and Recommendations

These results offer some pragmatic suggestions for researchers utilizing the SMS. First, in accordance with other researchers (John et al., 1996), it is recommended that the original 25-item SMS be used over the 18-item scale. This is primarily because the 18-item scale deletes the strongest item from the Other-Directness subscale, making a valid and reliable measure of this construct tenuous. Furthermore, the present analyses found the unrestricted models for all three subscales provided a superior fit to the data, indicating that the SMS subscale items are not equal in their relationship to their respective underlying construct. This calls into question the common practice of creating a summed score for each subscale. Instead, researchers may want to adopt a weighted scoring system such that items with stronger slopes are weighted

more heavily. Alternatively, researchers could omit the weakest items from the subscale before computing a summed score.

Finally, the results offer suggestions for the use of each subscale. For the Acting subscale, the thresholds all approximated zero. This implies that the subscale is useful for researchers interested in identifying two broad groups of individuals, those high and low in acting. However, if the researcher's goal is to make more finite distinctions, such as identifying those very high/low, then this subscale is unsuited for their task. Instead, one would need to create Acting subscale items that have greater variability in their threshold parameters and thus discriminate among varying levels of acting. For the Extraversion subscale, the thresholds were all negative. This implies that the subscale is only useful for researchers interested in identifying individuals low in extraversion. Only a small level of extraversion is required for endorsement of these items; therefore, they do not distinguish between moderate and high extraverts. If the researcher's goal is to separate those individuals who are low and high in extraversion, this subscale is not adequate for this purpose. Instead, one would need to create Extraversion subscale items that reflect the upper end of the continuum by having large positive threshold parameters. For the Other-Directedness subscale, the thresholds were mostly negative, with some extremely low thresholds. First, it is clear that none of the items are adequately assessing the other-directedness construct. Furthermore, several of the items are extremely poor and should possibly be omitted. For example, item 23 is the poorest item in the Other-Directedness subscale, but it is among the strongest items identified in the Extraversion subscale. It is therefore recommended that this item be removed from the Other-Directedness subscale and retained solely as an indication of extraversion. Also, more items should be developed that more strongly relate to other-directedness and that differentiate at the upper levels of this construct.

CONCLUSION

Item response theory is a useful psychometric tool and personality researchers can greatly benefit from IRT applications. Unlike more traditional approaches, IRT provides a detailed item-by-item analysis and allows researchers to identify when a particular scale is most useful and when it is an inappropriate application (for review of the use of IRT in personality, see Reise & Henson, 2003; Rouse, Finger & Butcher, 1999). The current IRT application provides useful information to anyone interested in studying the construct of self-monitoring. Armed with this information, researchers will be better able to understand what the SMS does and does not measure.

REFERENCES

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

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

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

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

Gangestad, S. W., & Snyder, M. (1985). To carve nature at its joints: On the existence of discrete classes in personality. *Psychological Review, 92,* 317-349.

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item-response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23,* 443-451.

Green, B., Bock, R., Humphreys. L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347-360.

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

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

John, O. P., Cheek, J. M., & Klohnen, E. C. (1996). On the nature of self-monitoring: Construct explication with Q-sort ratings. *Journal of Personality and Social Psychology, 4,* 763-776.

Muthén, L. K., & Muthén, B. O. (2000). *Mplus User's Guide*. Second Edition. Los Angeles, CA: Muthén & Muthén.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO-PI-R. *Assessment, 7,* 347-364.

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

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 *PSY*-5 scales. *Journal of Personality Assessment, 72*, 282-307.

Snyder, M., & Ickes, W. (1985). Personality and social behavior. In S. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol 3, pp. 883-946). New York: Random House.

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

Snyder, M., & Gangestad, S. (1986). On the nature of self-monitoring: Matters of assessment, matters of validity. *Journal of Personality and Social Psychology, 51,* 125-139.

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item and test scoring using item response theory.* Chicago: Scientific Software.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). *Multilog (version 7)* [Computer software]. Lincolnwood, IL: Scientific Software International.

Thissen, D., & Wainer, H. (Eds.) (2001). *Test scoring*. Mahwah, New Jersey: Lawrence Erlbaum.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer.

Applied Multivariate Research, Volume 13, No. 2, 2010, 145-159

THE FACETS OF JOB SATISFACTION: A NINE-NATION COMPARATIVE STUDY OF CONSTRUCT EQUIVALENCE

Catherine T. Kwantes [1]
University of Windsor

ABSTRACT

Archival data from an attitude survey of employees in a single multinational organization were used to examine the degree to which national culture affects the nature of job satisfaction. Responses from nine countries were compiled to create a benchmark against which nations could be individually compared. Factor analysis revealed four factors: Organizational Communication, Organizational Efficiency/Effectiveness, Organizational Support, and Personal Benefit. Comparisons of factor structures indicated that Organizational Communication exhibited the most construct equivalence, and Personal Benefit the least. The most satisfied employees were those from China, and the least satisfied from Brazil, consistent with previous findings that individuals in collectivistic nations report higher satisfaction. The research findings suggest that national cultural context exerts an effect on the nature of job satisfaction.

Keywords:

INTRODUCTION

With the rapid increase of globalization, national cultural context has become an important variable in understanding work behaviour. A current and critical challenge accompanying globalization in organizations is to understand how cultural values and expectations affect attitudes toward work (Ryan, Chan, & Ployhart, 1999; Erez, 1994). While job satisfaction may be a universal phenomenon, cultural factors may impact what employees actually focus on when determining what their level of job satisfaction is. Gelfand, Raver, and Ehrhart (2002) point out that that "a construct that is found to be universal may be manifested differently in different cultures" (p.218). Given that job satisfaction is related to the degree to which one"s needs related to work are satisfied (Shaffer, Joplin, Bell, Lau, & Oguz, 2000) as well as the fact that many of these needs are culturally defined (Olie, 1996) cultural determinants of job satisfaction deserve attention.

[1] Correspondence and requests for reprints should be addressed to Catherine T. Kwantes, Department of Psychology, University of Windsor, Windsor, Ontario, Canada, N9B 3P4, or email at ckwantes@uwindsor.ca.

Author Note: An earlier version of this paper was presented at the Academy of International Business annual conference, Monterey, CA, US, 2003.

While many researchers have examined the antecedents and consequences of job satisfaction, and others have compared overall levels of job satisfaction across groups (see, for example, Cass, Siu, Faragher, & Cooper, 2003; Judge, Heller, & Mount, 2002; Dormann & Zapf, 2001), little research has actually compared the components of job satisfaction across cultures. Theory suggests that a number of factors contribute to an individual‟s satisfaction with his/her job. However, the degree to which these factors actively make a difference in an individual‟s overall evaluation of the degree to which a job is satisfying, and what causes some factors to be salient in one context and others in another context, has received little attention. Rather, job satisfaction is frequently treated as either a dichotomous (one is satisfied or one is not satisfied with one‟s job), or ordinal (the degree to which one has overall job satisfaction) construct rather than examining the components of job satisfaction, and how those components are utilized by individuals in different contexts to determine satisfaction with his/her job (for example, Rode, 2004; Eskildsen, Kristensen, & Westlund, 2004).

Construct Equivalence Across Cultures

To understand global differences in job satisfaction, however, it is important to go beyond focussing on differences in overall estimates of job satisfaction alone; the transcultural equivalency of the construct of job satisfaction must be investigated. Therefore, one of the first, and most fundamental, questions in transcultural comparative research is the extent to which any given construct developed and originally measured in one culture can exist and operate similarly in another cultural context (see van de Vijver and Leung, 1997, Gelfand, Raver, and Ehrhart, 2002). Without establishing construct equivalence, it is impossible to have confidence that any differences found in cross-cultural comparative work are actually due to organizational phenomena rather than differences due to a methodological artefact. Ryan, Chan, and Ployhart (1999) recently addressed the issue of construct equivalence using an employee attitude survey in a multinational organization to examine language and culture as potential factors involved in measurement equivalence across multinational samples. They used samples in Australia, Mexico, Spain, and the U.S. Surveys were given in English to the Australian and American respondents, and in Spanish to the Mexican and Spanish respondents. Measurement equivalence, defined in this case as equal factor loadings across groups, was not found in this study. The researchers noted that a confirmatory factor analysis indicated a good fit, however, suggesting that the lack of measurement equivalence does not preclude equivalence in practical terms.

In a study that explicitly examined measurement equivalence of a job satisfaction measure across national contexts, Liu, Borg, and Spector (2004) examined the effects of both language and culture on the German Job Satisfaction Survey. They clustered the 15 countries and areas their sample came from into four cultural groups based on Schwartz‟s (1999) cultural model. Their findings indicate that measurement equivalence is particularly high when an instrument is used with groups where the language and culture are similar to the language and cultural context within which the instrument was developed.

Levels of Job Satisfaction Across Cultures

Evidence from recent transcultural research suggests that an employee‟s national culture context can affect his or her level of overall job satisfaction. For example, Van de Vliert and Janssen (2002) found international differences in job satisfaction level, as measured by a four-item measure derived from the International Value Research project among 42 different

countries. They found the highest level of job satisfaction was in Switzerland, followed by Norway and Iceland (with the percentage of respondents reporting favourable ratings for job satisfaction 85%, 78% and 76%, respectively) with the lowest levels in Hong Kong, Korea, and Taiwan (52%, 53%, and 55% favourable ratings, respectively) On the other hand, Sousa-Poza and Sousa-Poza (2000) examined data from a cross-national study of job satisfaction that spanned 21 countries, specifically looking at differences in job satisfaction, and found that, while differences in ranking do exist in employee reports of job satisfaction, their data indicated that overall job satisfaction was high across all samples. In order to explain the differences in relative job satisfaction levels, Sousa-Poza and Sousa-Poza examined the ratio of work inputs (such as schooling, work time, and degree of danger) with work outputs (such as income, security, advancement opportunities, intrinsic interest). Differences in this ratio were found to correlate with differences in the reported level of job satisfaction in each country. Further, some determinants of job satisfaction (interesting work, for example) were to be found in each of the countries in their sample, suggesting that job satisfaction is likely to be high, regardless of national culture context, when employees view their jobs as interesting, and where relationships with management are good. Other determinants (pay and job security, for example) were found in some countries and not in others, suggesting that some determinants of job satisfaction may be more salient within certain specific national contexts.

Determinants of Job Satisfaction

Another avenue of research examining job satisfaction across cultures has examined whether predictors of job satisfaction operate differently across cultural contexts. Sekaran (1981), for example, examined the degree to which measures of job satisfaction developed in the United States are appropriate for the Indian context. Identical questionnaires were given to bank employees in the U.S. and in India. Using factor analysis results, it was concluded that the job satisfaction measure was equally applicable in India and the United States. Several differences related to national culture context emerged, however, in the salience of different factors related to job satisfaction. Regression analysis indicated that job variety and stress were common predictors to both cultures, but income was an important predictor only in the U.S. and communication of policies was an important predictor only in India. In a study that compared more than two national contexts, Shaffer, Joplin, Bell, Lau, and Oguz (2000) examined the experiences of female employees with respect to the effect of gender discrimination on job satisfaction in the United States, Hong Kong, and the People‟s Republic of China (PRC). After establishing measurement equivalence through principal components factor analysis and structural equation modelling, they determined that job satisfaction was significantly higher in female employees from the United States than in Hong Kong, and the People‟s Republic of China, despite the fact that women in the United States reported more gender harassment than did women in the other countries. Female employees in Hong Kong and the PRC reported lower levels of job satisfaction, yet higher levels of gender evaluation and unwanted sexual attention respectively.

Other research has examined specific cultural factors and how they may impact an employee‟s job satisfaction. Eylon and Au (1999), for example, hypothesized that differences in job satisfaction would emerge between high and low power distance when employees were empowered by management. In a management simulation exercise they found that the interaction between cultural context and empowerment did affect work performance, however, contrary to their expectations, job satisfaction did not significantly differ between the two samples. Extending this idea somewhat, Chiu and Kosinski (1999) suggested that

personality traits are, to an extent, culturally determined, and that personality and culture may affect which factors are most salient to job satisfaction. Using samples from Singapore, Hong Kong, Australia, and the United States, they examined the interaction between personality traits and the cultural dimensions of individualism and collectivism, hypothesizing that individuals with high levels of individualism will also have high levels of positive affectivity and high levels of job satisfaction. Their results suggest that individualists with high positive affectivity and collectivists with high negative affectivity report the highest levels of job satisfaction.

Content Areas of Job Satisfaction

Job satisfaction is arguably one of the most researched concepts in organizational behaviour (Judge, Parker, Colbert, Heller, & Ilies, 2002) and this research has produced a wealth of constructs that potentially relate to job satisfaction. Intriguing evidence suggests, however, that these content areas of job satisfaction do not function equivalently across all cultural or national samples and, in fact, may not be perceived equally. In 1973, for example, O'Reilly and Roberts examined job satisfaction response patterns in samples of white and non-white nurses. They concluded that culture serves as a frame of reference when individuals think of job satisfaction, and therefore influences an employee's perceptions of their job. Furthermore, they suggest that culture may be a determinant in which factors are considered most salient to job satisfaction. Simonetti and Weitz (1972) found that nationality affected the degree to which aspects of a job were taken into account when individuals were asked to determine the overall degree of job satisfaction they experienced. In research that used employees located in three national contexts from a single multinational corporation, they found that patterns of factor contribution to overall job satisfaction differed in respondents in each location. Differences were also found in general levels of extrinsic and intrinsic job satisfaction dimensions. More recently, Al-Mashaan (2003) explored differences in Type A personalities and different content areas of job satisfaction in a sample of teachers in Kuwait and Egypt. No differences were found in overall level of job satisfaction between the two national samples, however, differences were found in several content areas of job satisfaction. Specifically, the Egyptian sample reported higher levels of satisfaction in areas such as satisfaction with work itself, satisfaction with organizational design and structure, and satisfaction with personal relationships. Thus, a more detailed examination of these areas of job satisfaction highlighted differences that were not immediately apparent in a more global comparison of the job satisfaction construct.

Present Research Goals

The goal of the present research was to assess similarities and differences in the content area of job satisfaction across a multinational sample as well as to examine the degree to which each factor makes a differential contribution to overall job satisfaction in each national context. The use of an attitude survey from a multinational organization in this research is particularly salient given the fact that these organizations often use instruments of this sort to collect data from employees. These instruments are characteristically developed within one cultural or national context, and then disseminated to the rest of the organization with the implicit assumption that the items and questions asked will have equivalent meaning to each employee, regardless of his or her national or cultural context. It is expected that the content areas of job satisfaction will not have construct equivalence across all national samples

(*Hypothesis 1)* as national context will result in different construct interpretations and meanings. In line with previous research, it is also expected that national samples will have differing levels of overall job satisfaction, despite the fact that each sample of employees came from the same multinational organization *(Hypothesis 2)*. Finally, it is expected that national context will affect the degree to which specific content areas predict employees" overall level of job satisfaction *(Hypothesis 3)* as cultural, economic, social, and political factors will differ in each setting.

METHOD

To examine job satisfaction across nine nations, several analyses were employed. The first analysis focussed on the determinants of job satisfaction reported by employees of a single organization, and the cross-cultural construct equivalency of these determinants by examining the degree to which agreement existed between nations in the factor structure of the data, as well as interpretation and perspectives on each factor. This was followed by an examination of the degree of job satisfaction reported by respondents in each nation to see if any statistical differences existed in the overall level of job satisfaction between nations. Finally, within nation analyses looked for any differential contribution of factors to explaining the level of job satisfaction.

Data came from an archival data set of employee attitudes measured in 1999 at a single organization headquartered in the United States with manufacturing and sales facilities located globally. The data were gathered as a result of an internal survey focussing on employee attitudes toward the organization and their supervisors. Respondents for this project were selected if they worked outside the United States, and were chosen from the larger data set based on the number of respondents from each country. Eliminating responses from countries with fewer than 150 respondents resulted in a sample of 6299. Participants were from England (n = 1471), Canada (n = 336), China (n = 1245), Romania (n = 907), Poland (n = 522), South Africa (n = 163), Brazil (n = 324), France (n = 838) and India (n = 493). Individuals responded to an internal organizational questionnaire asking them to report their perceptions of organizational activity ranging from the degree of communication received from supervisors to the degree of efficiency in organizational processes to the degree to which they benefited personally when the organization performed well. A Likert scale with responses ranging from 1 (strongly agree) to 5 (strongly disagree) was used.

RESULTS

Determination of Factor Structure. Questions regarding aspects of work that have been shown in previous theory or research (eg. Hackman & Oldham, 1976) to relate to job satisfaction were selected from the larger set of questions in the survey. Responses to these questions were then subjected to exploratory factor analysis to determine the underlying factor structure. In order to avoid any one national response set from having a disproportionate effect on the pooled data, the data were weighted so each nation contributed approximately 11.1% of the total sample. The weighted, pooled dataset was then factor analyzed using principal components analysis followed by varimax rotation. After eliminating factors with eigenvalues less than one, and after examining the scree plot, four factors emerged, contributing 60% of the total variance in job satisfaction, and were labelled Communication, Organizational Efficiency/Effectiveness, Organizational Support, and

Personal Benefit. Items were retained if their factor loading was greater than .40. Items with high loadings on two or more factors were eliminated from further analyses.

An item analysis was conducted, and items that reduced the total reliability of the factor solution were also eliminated from further analysis. One item was removed from the Organizational Efficiency/Effectiveness factor, and three from the Organizational Support factor. A final exploratory analysis was run, again using principal components analysis and varimax rotation, on the remaining items, resulting in the final scales. The Communication factor included seven items, such as "the performance expectations for my job have been clearly communicated to me." The Organizational Efficiency/Effectiveness factor also included seven items, such as "work flow processes are efficient and well organized where I work." The other two factors were comprised of two items after item analysis indicated that several items were not a good fit. The retained questions for Organizational Support included "I am provided with the support I need to do a quality job" while the Personal Benefit factor included "the better my performance the better my compensation will be." These factors are consistent with what has been found in previous research to relate to job satisfaction (for example: Bartlett, 2000; Pettit, Goris, & Vaught, 1997; Williams, Malos, & Palmer, 2002; Behson, 2002). Means, standard deviations, and reliabilities for each of the four factors, by nation, may be found in Table 1 and correlations between factors by country in Table 2.

Table 1
Means and Standard Deviations

	Organizational Communication			Organizational Efficiency /Effectiveness			Organizational Support			Personal Benefits		
	M	SD	α	M	SD	α	M	SD	α	M	SD	α
Brazil	2.48	.79	.90	2.86	.73	.74	2.36	1.06	.80	2.44	1.00	.51
Canada	2.52	.88	.90	3.01	.78	.84	2.28	1.16	.87	2.84	1.05	.77
China	2.15	.74	.86	2.28	.67	.85	2.39	1.08	.88	2.56	1.20	.76
England	2.82	.83	.89	2.95	.68	.81	2.52	1.04	.85	3.20	1.07	.76
France	2.41	.72	.87	2.58	.62	.79	2.32	0.89	.81	3.23	0.97	.70
India	2.61	.85	.83	2.70	.76	.79	2.88	1.20	.85	2.54	1.04	.56
Poland	2.22	.72	.90	2.38	.60	.81	2.44	0.94	.84	3.59	1.03	.87
Romania	2.19	.77	.85	2.35	.67	.75	2.51	1.18	.85	2.86	1.20	.79
South Africa	2.56	.84	.84	2.72	.67	.78	2.68	1.11	.70	2.94	1.07	.61

Exploration of Construct Equivalence. Before any meaningful comparison between responses across cultural divides can be made, it is necessarily to be sure that individuals in all cultures have similar understandings of the items asked, as well as the construct that the items are intended to measure (Liu, Borg, & Spector, 2004). Therefore, the second phase of this study examined the degree to which these four factors could be assumed to have the same meaning in each national cultural context. Following procedure recommended by van de Vijver and Leung (1997) for assessing construct equivalence, the factors from each nation were rotated using a Procrustes rotation, with the factor structure of the weighted, pooled sample serving as the target matrix. A coefficient of agreement between the factor loading on the weighted, pooled factor solution and the factor solution for each country was then

computed. This coefficient indicates the degree to which a factor has similar endorsement in two different samples.

Table 2
Correlations between Factors

		Efficiency	Support	Benefit
England	Communication	.597**	.481**	.484**
	Efficiency	1.00	.584**	.521**
	Support		1.00	.380**
Canada	Communication	.702**	.514**	.683**
	Efficiency	1.00	.573**	.625**
	Support		1.00	.370**
China	Communication	.544**	.329**	.481**
	Efficiency	1.00	.533**	.523**
	Support		1.00	.314**
Brazil	Communication	.570**	.424**	.388**
	Efficiency	1.00	.523**	.315**
	Support		1.00	.235**
France	Communication	.581**	.475**	.487**
	Efficiency	1.00	.548**	.461**
	Support		1.00	.395**
Romania	Communication	.611**	.508**	.470**
	Efficiency	1.00	.554**	.552**
	Support		1.00	.402**
Poland	Communication	.603**	.528**	.485**
	Efficiency	1.00	.573**	.478**
	Support		1.00	.413**
Romania	Communication	.534**	.479**	.499**
	Efficiency	1.00	.482**	.500**
	Support		1.00	.438**
South Africa	Communication	.386**	.479**	.304**
	Efficiency	1.00	.420**	.436**
	Support		1.00	.403**

*$p < .05$, **$p < .01$

Table 3
Coefficients of Agreement between National Sub-Sample and Pooled, Weighted Total Sample

	Communication	Efficiency	Support	Personal Benefit
Brazil	0.995	0.868	0.970	0.967
Canada	0.980	0.738	0.897	0.682
China	0.984	0.873	0.843	0.961
England	0.889	0.234	0.844	0.383
France	0.975	0.946	0.925	0.467
India	0.970	0.778	0.992	0.538
Poland	0.996	0.979	0.980	0.997
Romania	0.986	0.942	0.985	0.990
South Africa	0.986	0.927	0.890	0.868

In this case, each national sample was compared with the pooled, weighted data to assess the degree to which each nation is similar to or different from the total group on factor scores. The weighted, pooled data thus formed a norm against which individual national samples may be compared for similarity. A coefficient of agreement over .90 indicates a high degree of similarity on a particular factor between a specific national sample and the norm, while a coefficient over .95 indicates that there is virtually no difference between the factors in the two samples (van de Vijver & Leung).

Hypothesis 1 was supported, as no factors exhibited uniformly high construct equivalence, although some factors had more equivalence than others. Few differences were found on the Organizational Communication factor (see Table 3). Only England had a coefficient of agreement that indicated a less than identical construct equivalence with the norming sample (e = .88). On the Organizational Efficiency/Effectiveness factor, however, greater differences emerged. Only Poland (e = .98) had complete construct equivalence with the norming sample, although France, (e = .95), Romania (e = .94), and South Africa (e = .93) each had a high degree of similarity with the weighted, pooled data. Coefficients of agreement from all the other nations were considerably different, ranging from e = .87 (China) and e = .78 (India) to a low of e = .23 (England). Views of Organizational Support varied less widely, with coefficients of agreement ranging from e = .84 (China and England) to e = .99 (Romania and India). The fourth factor, personal benefit, resulted in the most variability. Brazil, China, Poland, and Romania all had coefficients of agreement greater than .90. On the other hand, vast differences from the norm were found for Canada (e = .68), England (e = .38), France (e = .47), India (e = .54) and South Africa (e = .87). These wide ranging differences suggest that the items on this factor are being interpreted very differently in different nations, and that the perspectives on these items similarly vary considerably by nation.

Comparison of Overall Job Satisfaction Level. A single question was asked regarding the degree to which individuals were satisfied with their job. While single item indices are generally not considered sufficient to measure a construct, job satisfaction may be an exception to this heuristic. Wanous, Reichers, & Hudy (1997), for example, suggest that the use of a single item to measure job satisfaction is not a "fatal flaw" in the research process. There is support for the idea that single-item measures are sufficient when the construct is a narrow one or is unambiguous to the respondent (Sackett & Larson, 1990) and some research suggests that in the case of measuring job satisfaction, a single-item measure may actually be

preferable to a scale of components of job satisfaction (Nagy, 2002; Scarpello & Campbell, 1983). Accordingly, a one-way ANOVA was performed with nation as the independent variable and with responses to the single item job satisfaction question serving as the dependent variable, followed by multiple comparisons using a Bonferroni adjustment as well as a Student-Newman-Kuhls post hoc analysis. Hypothesis 2 was supported, as the results indicated significant differences, with China responding with the highest degree of job satisfaction (M = 3.66), followed by England (M = 2.85), then India (M = 2.71). Five countries comprised the next group of respondents, with no statistically significant difference between them: S. Africa (M = 2.55), France (M = 2.51), Canada (M = 2.50), Romania (M = 2.50) and Poland (M = 2.41). Respondents from Brazil reported significantly lower job satisfaction then those from other nations (M = 2.10). This finding is somewhat consistent with earlier research indicating that, in general, individuals from more collectivistic countries are more likely to be satisfied with their jobs than individuals from individualistic countries (Hui, Yee, & Eastman, 1995) as Brazil is considered to be moderately individualistic while China is considered a more collectivistic nation (Hofstede, 2001).

Comparison of Job Satisfaction Content Areas. To compare constructs across different contexts in cross-cultural research it is not unusual to construct one hierarchical regression equation that includes nation in the first step, then examines the degree to which independent variables have the ability to predict the dependent variable and look for interaction effects with nation and predictor variables. However, in this case, it was felt that this would be inappropriate given the fact that construct equivalence was not found for all variables in all samples. Accordingly, only within country analyses were performed to find the degree to which each of these factors significantly predicted job satisfaction within each individual group. Job satisfaction was regressed on Communication, Organizational Efficiency/Effectiveness, Organizational Support, and Personal Benefit. The regression equation significantly predicted job satisfaction in all nine countries, with the amount of variance explained ranging from a high of 55% in Canada to a low of 17% in China (See Table 4).

Table 4
Summary of Regression Analysis for Factors Predicting Overall Job Satisfaction

			Communication		Efficiency		Support		Personal Benefit	
	R^2	Adj R^2	B	SE	B	SE	B	SE	B	SE
Brazil	0.32**	0.31**	0.13	.07	0.16*	.08	0.30**	.05	0.13**	.05
Canada	0.55**	0.54**	0.22**	.07	0.45**	.08	0.24**	.05	0.01*	.05
China	0.17**	0.17**	0.16**	.05	0.11	.06	0.19**	.03	0.14**	.03
England	0.38**	0.38**	0.30**	.03	0.25**	.04	0.16**	.03	0.17**	.03
France	0.44**	0.44**	0.19**	.04	0.22**	.05	0.25**	.03	0.17**	.03
India	0.50**	0.50**	0.11*	.06	0.36**	.07	0.38**	.04	0.10*	.04
Poland	0.39**	0.38**	0.12	.05	0.35**	.07	0.11**	.04	0.20**	.03
Romania	0.44**	0.44**	0.17**	.04	0.14**	.05	0.22**	.03	0.23**	.03
South Africa	0.28**	0.26**	0.42**	.09	-0.01*	.12	0.12	.08	0.20*	.07

While the equation predicting job satisfaction was significant in all countries, the degree to which each of the four factors contributed to the explanation of job satisfaction in different countries differed, supporting Hypothesis 3. All four factors were significant predictors in England, Canada, France, and Poland. Communication did not provide an independent

contribution to the job satisfaction regression equation in India and Brazil, while Efficiency did not in China. The regression equation for South Africa included only communication and personal benefit as significant predictors of job satisfaction.

DISCUSSION

The results from these analyses suggest that, while differences exist in the level of job satisfaction in different countries, even greater discrepancies exist in defining what aspects of a job contribute to a perception of satisfaction or dissatisfaction. Attention in transcultural research has shifted from investigating what factors are important in defining satisfaction with the job to examining how antecedents differentially predict overall job satisfaction in different contexts (for example, Vigoda, 2001). The findings of this research suggest that it would be prudent to return to an examination of what aspects of a job are valued in different cultures and how those value differences are reflected in how aspects of a job are perceived.

Organizational Communication and Organizational Support were the components of job satisfaction where most agreement existed across national samples, while Organizational Efficiency/Effectiveness and Personal Benefit resulted in much more disparate perspectives. Of these four factors, the convergence of understanding with respect to the Organizational Communication construct and the divergence of understanding with respect to the construct of Personal Benefit are the most easily understood. The finding that so much agreement existed with regard to communication may be due to the fact that most of the items relating to this factor refer to behaviours and practices that are common in organizations globally, and therefore would likely have similar meaning to individuals across national borders. Further, the fact that all respondents were from one organization may have reduced the degree to which disparate understandings or perspectives exist on this factor.

On the other hand, the items regarding personal benefit were ones that allowed more room for unique interpretations. The term "benefit" may be thought of in several ways, for example, as a monetary benefit or as a status benefit. In this survey, no attempt was made to clarify specific interpretations of this term, therefore it is possible that any agreement or disagreement that exists between samples is the result of wording effects rather than actual agreement or disagreement about a unitary construct.

National culture orientation, however, may explain some of the differences found in the other dimensions. Cultural expectations may vary considerably with regard to what does and what does not constitute organizational support. Hofstede (1997) points out that organizations are focussed on individuals in some societies and on groups in others. Expected support in an individualist context is likely to include support aimed at individual effort, while in a collectivistic context, support is more likely to be defined as interventions aimed at a group effort. Yoon and Lim (1999) found support for the idea that perceptions of what constitutes organizational support differ as the result of individualism and collectivism. In a collectivistic context, relationships are perceived as an integral part of organizational support, while in more individualistic contexts, perceptions of support have more of an economic exchange component. Similarly, in a high power distance context, organizational support is more likely to be perceived as clear and explicit direction while in a low power distance context, support is more likely to be seen as providing a climate where those closest to the work are supported to make their own decisions relating to the work process. These cultural differences in perspective likely spill over into the differences in perspective regarding organizational efficiency/effectiveness as well, since the two concepts are highly interrelated (Yoon & Lim, 1999). Indeed, empirical evidence supports the idea that national

culture value orientations do affect work attitudes in general, and job satisfaction in particular (Sparrow & Wu, 1998; Bae & Chung, 1997; Saiyadain, 1985).

While the results of this study do not explicitly tease out culture from other national differences that may affect attitudes toward job satisfaction, some intriguing findings emerged. Respondents from both Brazil and India indicated that organizational communication was not a salient factor in determining job satisfaction. Both of these nations have been found to have hierarchical social cultures – Hofstede (2002) found both Brazil and India to rank 10th and 14th respectively out of 50 countries on power distance, with Power Distance Index Values almost twice that of Great Britain or Canada. Endorsement of power distance implies that consultative managers are viewed with less respect than authoritative managers in organizations, and that task oriented leadership is preferred to relationship oriented leadership (Hofstede, 2002). Not expecting organizational communication in high power distance contexts corresponds to these characteristics, and aids in explaining why this factor did not contribute to explaining the degree of job satisfaction in these national contexts.

Respondents from China indicated that only organizational efficiency/effectiveness was unrelated to job satisfaction. Again, national culture values may be seen at work here. Organizational cultures in the Chinese context have been shown to distinctly reflect Chinese social culture (Kwantes, Boglarsky, & Kuo, 2004). Imbedded in Chinese culture is a strong desire for social acceptance (Bond, 1996; Hosfstede, 2001), and therefore a common focus is on the preservation of social harmony – a focus which may, at times, be at odds with the concepts of efficiency and effectiveness. Furthermore, China consistently ranks as a culture with a very strong emphasis on long-term orientation with its attendant characteristics of persistence, perseverance, and a perspective that precludes an expectation of quick results (Hofstede, 2002). Discipline, moderation, tolerance, harmony, and non-competitiveness are values that have been shown to be consistent with a long-term orientation. Examining the individual items on the Efficiancy/Effectiveness factor, it is possible that the way this factor was operationalized in this study may have led to weak endorsement of these items in the Chinese cultural context, as the items emphasized encouragement to take calculated risks, permission to make one‟s own decisions to improve organizational effectiveness, and efficient processes.

In the sample from South Africa, only organizational communication and personal benefits were considered salient to job satisfaction. South Africa has ranked relatively high on individualism in previous studies (Hofstede, 2002). This individualism, combined with the uncertainty, economic and political turmoil prevalent in South Africa today may aid in explaining these results. Organizational communication provides some clarity for employees regarding both individual and organizational progress, and this in turn provides some information regarding the viability of the organization. In an uncertain time, this information, along with the degree to which one personally benefits from one‟s job become more tangible than support or efficiency of the organization as a whole. In a national context of individualism, it is not surprising that factors relating most directly to an individual‟s personal outcomes would be most salient.

CONCLUSION

Satisfaction is one of the most widely used attitudinal measures in organizations, and is also one of the most widely used attitudinal variables in research. It is frequently used as a summary measure of other employee attitudes such as acceptance and contentment (Hodson, 2002). From the perspective of both research and practice, a better understanding of how

national culture contexts affect both the definition and content areas of job satisfaction is warranted. This research contributes to the literature by providing initial evidence that the operationalization of job satisfaction may be affected by national context, and that any comparison of job satisfaction across national or cultural boundaries must take measurement equivalence but also construct equivalence into account. This is particularly salient in multinational corporations where comparisons across national contexts are used for organizational development decision making.

It must be noted, however, that this research focussed on national context and not explicitly on cultural context - while national context differences explained 18.5% of the variance in job satisfaction in this research project, culture alone does not likely explain all national differences. Differences in job attitudes may be the result of national political economy variables, national labor markets (Sparrow & Wu, 1997) and other contextual factors unrelated to cultural values. Organizational factors such as HR policies and practices may also be involved in attitudinal differences. While examining job satisfaction within the confines of a single organization reduces the variability due to organizational factors, it does not eliminate it.

Nevertheless, the findings still strongly suggest a renewed investigation into the construct and content area equivalency of job satisfaction in different national culture contexts. The data used were archival and from an internal organizational attitude survey, therefore not chosen a priori to explicitly address this research question. Future research using both qualitative and quantitative methods would be better suited to explicate possible differences in job satisfaction across national cultures. Isolating the research project to focus on a few number of countries would allow for a finer grained approach to hypothesizing the direction of differences, but more importantly, the potential causes for these differences. Future research including more variables, for example, explicit cultural value orientations of the respondents and demographic variables would help to elucidate the extent to which cultural factors provide independent explanations for attitudinal differences or act in conjunction with other variables and provide a much more detailed understanding of the meaning of the construct in different contexts.

REFERENCES

Al-Mashaan, O. S. (2003). Comparison between Kuwaiti and Egyptian teachers in Type A behavior and job satisfaction: A cross-cultural study. *Social Behavior and Personality, 31,*(5), 523-524.

Auster, E. R. (2001). Professional women‟s midcareer satisfaction: Toward an explanatory framework. *Sex Roles, 44,* 11-12, 719-750.

Bae, K., & Chung, C. (1997). Cultural values and work attitudes of Korean industrial workers in comparison with those of the United States and Japan. *Work and Occupations, 24,* 1, 80-96.

Bartlett, C. (2000). Supervisory communication and subordinate job satisfaction: The relationship between superiors‟ self-disclosure, offers of help, offers of cooperation, frequency of contact, trust and subordinates‟ job satisfaction. *Public Library Quarterly, 18,* 1, 9 – 30.

Behson, S. J. (2002). Which dominates? The relative importance of work-family organizational support and general organizational context on employee outcomes. *Journal of Vocational Behavior, 61,* 1, 53-72.

Bond, M. H. (1996). Chinese values. In M. H. Bond (Ed.). *The Handbook of Chinese Psychology* (pp. 208-226). NY: Oxford University Press.

Cass, M. H., Siu, O. L., Faragher, E. B., & Cooper, C. L. (2003). A meta-analysis of the relationship between job satisfaction and mployee health in Hong Kong. *Stress & Health, 19,* 2, 79-95.

Chinese Culture Connection. (1987). Chinese values and the search for culture-free dimensions of culture. *Journal of Cross-Cultural Psychology, 18,* 143-164.

Chiu, R. K., Kosinski, F. A. (1999). The role of affective dispositions in job satisfaction and work strain: Comparing collectivist and individualist societies. *International Journal of Psychology, 34,* 1, 19-28.

Dormann, C., & Zapf, D. (2001). Job satisfaction: A meta-analysis of stabilities. *Journal of Organizational Behavior, 22,*5, 483-504.

Erez, M. (1994). Toward a model of cross-cultural industrial and organizational psychology. In H. C. Triandis, M. D. Dunnete, & L. M. Hough (Eds). *Handbook of industrial and organizational psychology, Vol. 4.* pp. 559-608. Palo Alto, CA, US: Consulting Psychologists Press.

Eskildsen, J. K., Kristensen, K., & Westlund, A. H. (2004). Work motivation and job satisfaction in the Nordic countries. *Employee Relations, 26,* 2, 122-136.

Eylon, D., & Au, K. Y. (1999). Exploring empowerment cross-cultural differences along the power distance dimension. *International Journal of Intercultural Relations, 23,* 3, 373-385.

Gelfand, M. J., River, J. L., & Ehrhart, K. H. (2002). Methodological issues in cross-cultural organizational research. In S. Rogelberg (Ed.) *Handbook of Research Methods in Industrial and Organizational Psychology.* Malden, MA, US: Blackwell.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior & Human Decision Processes, 16,* 2, 250-279.

Hodson, R. (2002). Demography or respect? Work group demography or organizational dynamics as determinants of meaning and satisfaction at work. *British Journal of Sociology, 53,* 2, 291-317.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage.

Hofstede. G. (1997). *Cultures and Organizations: Software of the Mind.* New York: McGraw-Hill.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Hui, C. H., Yee, C., & Eastman, K. L. (1995). The relationship between individualism-collectivism and job satisfaction. *Applied Psychology: An International Review, 44,* 3, 276-282.

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87,* 3, 530-541.

Judge, T. A., Parker, S. K., Colbert, A. E., Heller, D., & Ilies, R. (2002). Job satisfaction: A cross-cultural review. In N. Anderson & D. Ones (Eds.) *Handbook of Industrial, Work and Organizational Psychology, Vol 2. Organizational Psychology*. 25-52,Thousand Oaks, CA, US: Sage Publications.

Kwantes, C. T., Boglarsky, C. A., & Kuo, B. C. H. (2004). One organization, three nations: Harmonies in organizational culture. Paper presented at the Third Biennial Conference on Intercultural Research, Taipei, Taiwan.

Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German Job Satisfaction Survey used in a multinational organization: Implications of Schwartz‟s culture model. *Journal of Applied Psychology, 88*(6), 1070-1082.

Olie, R. (1996). The „culture‟ factor in personnel and organizational policies. In A. W. Harzing & J. Van Ruysseveldt (Eds.) *International Human Resource Management* (pp. 124-133). London: Sage.

O‟Reilly, C. A., & Roberts, K. H. (1973). Job satisfaction among whites and nonwhites: A cross-cultural approach. *Journal of Applied Psychology, 57,* 3, 295-299.

Pettit, J. D., Jr., Goris, J. R., & Vaught, B. C. (1997). An examination of organizational communication as a moderator of the relationship between job performance and job satisfaction. *Journal of Business Communication, 34,* 1, 82-98.

Redding, G. & Wong, G. Y. Y. (1986). The psychology of Chinese organizational behavior. In M. H. Bond (Ed.), *The psychology of the Chinese people* (pp. 267-295). NY: Oxford University Press.

Rode, J. C. (2004). Job satisfaction and life satisfaction revisited: A longitudinal test of an integrated model. *Human Relations, 57,* 9, 1205-1230.

Ryan, A. M., Chan, D., & Ployhart, R. E. (1999). Employee attitude surveys in a multinational organization: considering language and culture in assessing measurement equivalence. *Personnel Psychology, 52,* 1, 37-58.

Sackett, P. R., & Larson, J. R., Jr. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough, Eds. *Handbook of Industrial and Organizational Psychology, Vol. 1, 2nd Ed.* P. 419-489. Palo Alto, CA, US: Consulting Psychologists Press, Inc.

Saiyadain, M. S. (1985). Personal characteristics and job satisfaction: India-Nigeria comparison. *International Journal of Psychology* Special Issue: Job satisfaction, *20,* 2, 143-153.

Scarpello, V., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology, 36,* 3, 577-600.

Sekaran, U. (1981). Are U.S. organizational concepts and measures transferable to another culture? An empirical investigation. *The Academy of Management Journal, 24,* 2, 409-417.

Shaffer, M. A., Joplin, J.R.W., Bell, M. P., Lau, T., & Oguz, C. (2000). Gender discrimination and job-related outcomes: A cross-cultural comparison of working women in the United States and China. *Journal of Vocational Behavior, 57,* 3, 395-427.

Simonetti, S. H., & Weitz, J. (1972). Job satisfaction: Some cross-cultural effects. *Personnel Psychology, 25,* 1, 107-118.

Sousa-Poza, A., & Sousa-Poza, A. A. (2000). Well-being at work: A cross-national analysis of the levels and determinants of job satisfaction. *Journal of Socio-Economics, 29,* 517-538.

Sparrow, P., & Wu, P-C. (1998). Does national culture really matter? Predicting HRM preferences of Taiwanese employees. *Employee Relations, 20,* 1, 26-56.

Van de Vijver, F. J. R., & Leung, K. (1997). Methods and Data Analysis of Comparative Research. In J. W. berry, Y. H. Poortinga, & J. Pandey (Eds). *Handbook of Cross-Cultural Psychology: Volume 1 Theory and Method.* Boston: Allyn and Bacon.

Van de Vliert, E., & Janssen, O. (2002). Better than: performance motives as roots of satisfaction across more and less developed countries. *Journal of Cross-Cultural Psychology, 33,* 4, 380-397.

Vigoda, E. (2001). Reactions to organizational politics: A cross-cultural examination in Israel and Britain. *Human Relations, 54,* 11, 1483-1518.

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82,* 2, 247-252.

Williams, M. L., Malos, S. B., & Palmer, D. K. (2002). Benefit system and benefit level satisfaction: An expanded model of antecedents and consequences. *Journal of Management, 28,* 2, 195-215.

Yoon, J., & Lim, J-C. (1999). Organizational support in the workplace: The case of Korean hospital employees. *Human Relations, 52,* 7, 923-945.

Applied Multivariate Research, Volume 13, No. 2, 2010, 161-174

AN INTERPERSONAL CIRCUMPLEX/FIVE-FACTOR MODEL ANALYSIS OF THE EATING DISORDERS INVENTORY-3

Jeffrey B. Brookings [1]
Wittenberg University

Corey D. Beilstein
Seed Strategy, Inc.

ABSTRACT

A combined interpersonal circumplex/five-factor model approach was used to investigate personality correlates of Eating Disorders Inventory-3 (EDI-3; Garner, 2004) scales for a non-clinical sample of 234 college women. EDI-3 non-symptom scales and composites had appreciable loadings in the two-dimensional interpersonal circumplex space, with angular locations ranging mainly from Cold (180°) to Submissive (270°). In the five-factor analyses, Neuroticism made significant positive contributions to all of the EDI-3 scales and composites; Conscientiousness made contributions (all negative, save one) to 11 of the 18 scales. The results affirm the centrality of negative affect (i.e., Neuroticism) in disordered eating, but highlight also the importance of assessing interpersonal deficits, which in previous studies have been associated both with the etiology of eating-related problems and increased risk of dropout from treatment. Finally, collapsing or "weighting" EDI-3 item scores may compromise unnecessarily the psychometric properties of the scales—particularly in non-clinical populations—and we recommend derivation of additional EDI-3 norms, based on unweighted item scores.

Keywords: *Eating Disorders; Interpersonal Circumplex; Big Five*

INTRODUCTION

The Eating Disorders Inventory (EDI), now in its third edition (EDI-3; Garner, 2004), is a "…self-report measure of psychological traits or constructs shown to be clinically relevant in individuals with eating disorders" (p. 1). The EDI's popularity for clinical assessment and treatment has grown steadily since its publication in 1983 (Theander, 2004), and researchers are increasingly using the EDI to evaluate theoretical models of the structure, antecedents, and consequences of eating disorders. Even though the original EDI was created specifically to evaluate a specific model of anorexia nervosa (Nylander, 1971), its breadth of content (the EDI-3 has 12 scales and 6 composites) makes it attractive to researchers investigating a variety of theoretical perspectives on eating disorders. The EDI-3 composites, for example,

[1] Jeffrey B. Brookings, Ph.D., Department of Psychology, Wittenberg University, P.O. Box 720, Springfield, OH 45501, Email: jbrookings@wittenberg.edu
Note: We thank Stephanie Little for comments on an earlier draft of this article.

assess problems ranging from weight preoccupation to emotional dysregulation to general psychopathology.

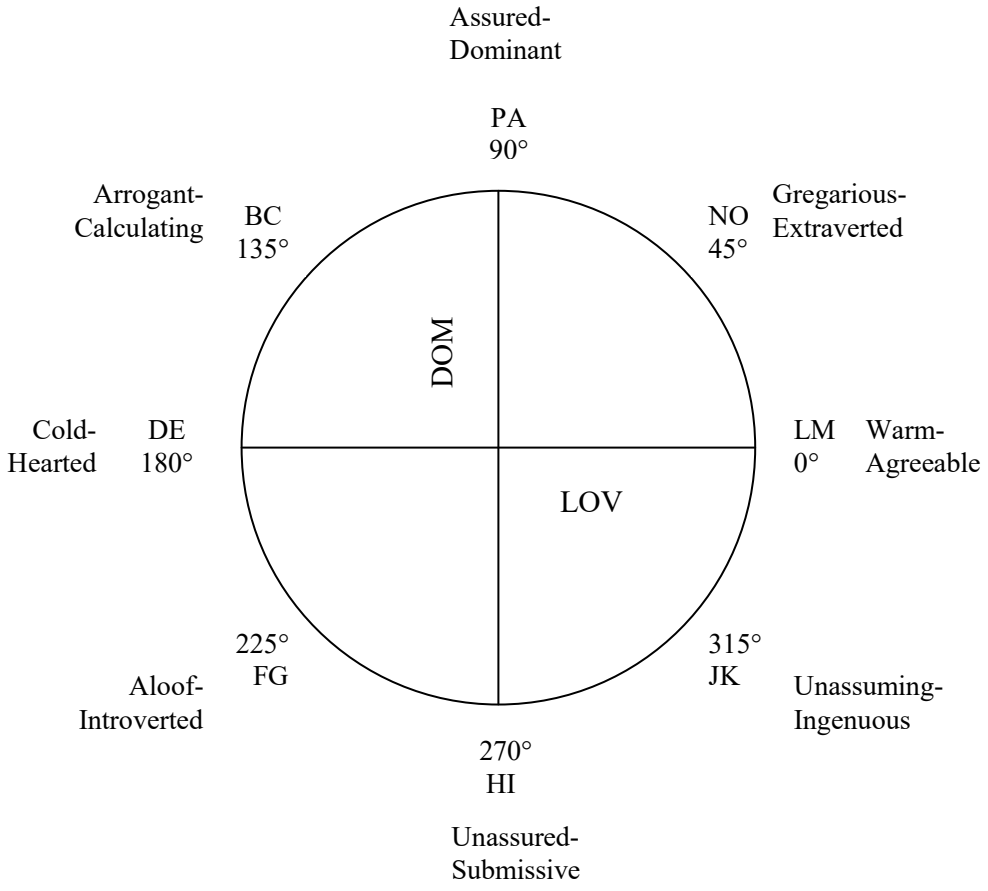
Only three of the 12 EDI-3 scales contain eating-related content; the remaining nine scales measure constructs that covary with eating disorders, including personality traits (e.g., low self-esteem, perfectionism). Among these, interpersonally-oriented personality traits have received relatively little attention in eating disorder research, despite evidence linking them to self-reports of disordered eating (e.g., Madison, 1997), and despite the emergence of interpersonal approaches to assessing and treating eating disorders, which view eating-related symptoms as pathological responses to interpersonal problems (Wilfley, Stein, & Welch, 2005) or as self-regulation failure triggered in part by interpersonal stressors (Horowitz, 2004). The assessment of interpersonal constructs—via patients' self-reports of their interpersonal relationships—may thus fill gaps in the diagnostic picture offered by other approaches (e.g., cognitive-behavioral). In addition, clearer delineation of the interpersonal features of the EDI scales in non-clinical populations may identify interpersonal traits associated with emerging/subclinical eating problems, which occur frequently among college women (Mintz & Betz, 1988). From its inception, the EDI has included items with interpersonal content, and the most recent version, the EDI-3 (Garner, 2004), includes an Interpersonal Problems Composite. However, the only published study of interpersonal correlates of the EDI among non-patients (Brookings & Wilson, 1994) used an earlier version of the scale, and the interpersonal scales used in that study precluded fine-grained interpersonal analyses.

The Interpersonal Circumplex

The interpersonal circumplex has a long research history in clinical assessment (Benjamin, 1996a; Horowitz, 2004; Leary, 1957) and construct validation (Gurtman, 1992; Wiggins & Broughton, 1985). Briefly, circumplex models of personality view interpersonal traits as blends of two orthogonal, superordinate constructs, Dominance (DOM) and Love (LOV) (Wiggins & Trapnell, 1996). In numerous studies (see Gurtman & Pincus, 2003), it has been demonstrated that the pattern of correlations among carefully-constructed interpersonal trait scales conforms closely to a circular or *circumplex* (Guttman, 1954) ordering about DOM and LOV (see Figure 1), and that the resulting two-dimensional model provides a succinct but comprehensive representation of the interpersonal trait domain. The most visible products of circumplex research are scales for assessing interpersonal traits and problems, including the Interpersonal Adjective Scales (IAS; Wiggins, 1995), the Inventory of Interpersonal Problems-Circumplex (Horowitz, Alden, Wiggins, & Pincus, 2000), and the Structural Analysis of Social Behavior (Benjamin, 1996b).

Methods for the second application of the interpersonal circumplex, construct validation, were introduced by Wiggins and Broughton (1985) and elaborated subsequently by Gurtman and others (e.g., Gurtman, 1992, 1994; Gurtman & Pincus, 2003; Tracey, 2000). In this application, measures of target constructs (e.g., eating disorder scales) are evaluated with respect to the extent and quality of their interpersonal content or "interpersonalness" (Gurtman, 1991), estimated from the vector length (extent) and angular placement (quality) of a measured variable—item, scale, composite—in the two-dimensional circumplex space. Vector length reflects a variable's level of "interpersonal saturation" (i.e., longer vector = more interpersonal content), whereas angular placement provides substantive information about the variables' interpersonal content.

Figure 1. **The interpersonal circumplex.**



The Five-Factor Model of Personality

Factor analytic studies (e.g., McCrae & Costa, 1989) provide evidence that DOM and LOV are rotational variants of Extraversion and Agreeableness, respectively, the interpersonal dimensions in the five-factor model of personality (McCrae & John, 1992). Specifically, DOM has been characterized as "cold" Extraversion and LOV as "bold" Agreeableness (Trapnell & Campbell, 1999). However, the most frequently used interpersonal circumplex measure, Wiggins' (1995) IAS, contains no items assessing the remaining five-factor model traits: Neuroticism, Conscientiousness, and Openness to Experience. To meet the needs of researchers for a combined interpersonal circumplex/five-factor model measure, Trapnell and Wiggins (1990) added items assessing these three traits to the IAS. The resulting measure, the Interpersonal Adjective Scales Revised-Big Five (IASR-B5), was used in this study.

The Current Research

In a study of the interpersonal profiles of patients with eating disorders, Madison (1997) identified two prominent subgroups: One had a cold/hostile interpersonal style but the other group was located—surprisingly—in the agreeable/extraverted portion of the circumplex. In a non-clinical sample of female undergraduates (Brookings & Wilson, 1994), NEO Personality Inventory (NEO-PI; Costa & McCrae, 1985) Extraversion correlated positively with EDI (Garner & Olmsted, 1984) Drive for Thinness and negatively with Interpersonal Distrust, and Agreeableness correlated negatively with Ineffectiveness and Interpersonal Distrust. The cross-sectional designs of these studies precluded inferences about the causal priority of interpersonal difficulties and disordered eating. As a practical matter, though, a systematic interpersonal analysis of the EDI could identify scales or combinations of scales with high levels of interpersonal content. These constructs, in turn, could be investigated as risk factors for eating disorders in non-clinical female populations (e.g., college women), for whom "subthreshold" dieting and binging are disturbingly frequent (Kurth, Krahn, Nairn, & Drewnowski, 1995).

Brookings and Wilson's (1994) interpersonal analyses of the EDI scales were limited by the original NEO-PI, which included only a global measure of Agreeableness (LOV). Also, the recently published EDI-3 (Garner, 2004) reflects—in terms of its scale and composite scale structure—a more distinct interpersonal emphasis than did the original EDI. The objective of this investigation, then, was a combined interpersonal circumplex/five-factor model analysis of the EDI-3 scales and composites for a non-clinical female sample. Based on previous empirical and clinical findings (Brookings & Wilson, 1994; Madison, 1997; Wilfley, et al., 2005), we predicted that: a) there would be substantial interpersonal loadings for EDI-3 scales and composites related to interpersonal problems and general psychological maladjustment, but not the eating disorder symptom scales (Bulimia, Drive for Thinness, Body Dissatisfaction); b) angular locations for the non-symptom scales would vary widely, reflecting the mixed findings reported in studies of eating disorders and extraversion; and c) across all EDI-3 scales and composites, five-factor model Neuroticism would explain the most unique variance.

METHOD

Participants

The participants were 234 female undergraduates at a small Midwestern university, who participated in the study to earn extra credit in introductory psychology courses. The mean age of the women was 20.4 (*SD* = 4.2).

Measures

EDI-3. The participants in this study actually completed the EDI-2 (Garner, 1991), as data collection occurred just prior to publication of the EDI-3 (Garner, 2004). However, because the 91 items comprising the EDI-2 and EDI-3 are identical, our participants' responses were scored for both sets of scales. Differences between the two versions are in the number of scales (11 for the EDI-2 and 12 for the EDI-3), minor differences in the allocation of items to the scales and, on the EDI-3, addition of the following six composite scales: Eating Disorder Risk (Drive for Thinness + Bulimia + Body Dissatisfaction); Ineffectiveness (Low Self-

Esteem + Personal Alienation); Interpersonal Problems (Interpersonal Insecurity + Interpersonal Alienation); Affective Problems (Interoceptive Deficits + Emotional Dysregulation); Overcontrol (Perfectionism + Asceticism); and General Psychological Maladjustment (sum of the nine non-symptom scales).

There is one additional difference, in item scoring: For the EDI-2, the six-point item responses (*Always* to *Never*) are "weighted" as follows: *Always* = 3, *Usually* = 2, *Often* = 1; *Sometimes*, *Rarely*, and *Never* are all assigned scores of 0. On the EDI-3, the item weightings are *Always* = 4, *Usually* = 3, *Often* = 2, *Sometimes* = 1; *Rarely* and *Never* are scored 0. In the EDI-2 manual, Garner (1991) advances a "…rational-theoretical rather than empirical…" (p. 6) justification for weighted item scoring. Specifically, he argued that scores on the non-symptomatic portion of the continuum should not contribute to total scores on a scale designed to assess psychopathology. There is evidence, however, that the variability lost by truncating item scores compromises the sensitivity of the scales, particularly in non-clinical samples (Schoemaker, Van Strien, & Van der Staak, 1994; Van Strien & Ouwens, 2003). To allow comparison of our data with those from other studies, we report descriptive statistics for both the EDI-2 and EDI-3 scales, using weighted and unweighted item scores. The presentation of interpersonal statistics follows the same format. However, to conserve space and simplify the presentation of our findings, circumplex plots and results of the five-factor model multiple regression analyses are presented only for the EDI-3 scales and composites, computed from unweighted item scores.

IASR-B5. The IASR-B5 (Trapnell & Wiggins, 1990) is a 124-item, interpersonal circumplex/five-factor model measure. The circumplex portion is composed of the 64 items from the IAS (Wiggins, 1995), which are scored for eight octant scales and the orthogonal superordinate constructs DOM and LOV (see Figure 1). To these 64 items, Trapnell and Wiggins (1990) added 20 adjective markers each for Neuroticism, Conscientiousness, and Openness to Experience. Respondents rate how accurately each item describes them on a scale ranging from 1 (*Extremely Inaccurate*) to 8 (*Extremely Accurate*). Evidence for the construct validity of the IASR-B5 is summarized by Wiggins and Trobst (2002).

Data Analyses

Several methods are available to calculate projections of variables in a circumplex space (see Gurtman & Pincus, 2003), all of which begin with assessment of the extent to which the correlational pattern among the first-order scales (i.e., the IASR-B5 octant scales) conforms to the expected circular ordering. For the analyses reported here, which are based on the vector method described by Wiggins and Broughton (1985, 1991), we then calculated projections of the EDI-3 scales and composites in the two-dimensional circumplex space. This was accomplished using the following formulas: Vector length (VL) = $(DOM^2 + LOV^2)^{½}$ and angular placement (AP) = arctangent (DOM/LOV), where DOM and LOV are correlations of the EDI-3 variables with scores on the respective interpersonal constructs. For the five-factor model analyses, simultaneous multiple regression analyses estimated the individual and collective contributions of the five-factor domain scales to each EDI-3 scale and composite.

RESULTS

Descriptive Statistics

Descriptive statistics for the EDI-2 and EDI-3 measures are presented in Table 1. Twelve percent of our sample had EDI-2 Drive for Thinness scores at or above the EDI-2 cutoff for

Table 1

Descriptive Statistics for the EDI-2 Scales, EDI-3 Scales, and EDI-3 Composites

	Weighted Scores			Unweighted Scores		
Measure	M	SD	α	M	SD	α
EDI-2 Scales						
Drive for Thinness (7)	5.2	5.8	.90	22.5	8.6	.91
Bulimia (7)	1.8	3.2	.82	15.5	6.4	.86
Body Dissatisfaction (9)	10.8	8.4	.93	34.7	11.2	.92
Ineffectiveness (10)	2.5	3.5	.80	24.6	7.8	.88
Perfectionism (6)	6.6	4.1	.72	22.6	5.6	.75
Interpersonal Distrust (7)	2.1	2.6	.69	17.6	5.2	.77
Interoceptive Awareness (10)	2.8	4.1	.81	25.6	7.8	.86
Maturity Fears (8)	3.3	3.4	.75	23.2	6.1	.81
Asceticism (8)	3.8	2.9	.52	21.8	4.9	.60
Impulse Regulation (11)	2.2	3.5	.76	25.0	6.7	.79
Social Insecurity (8)	2.7	2.6	.65	20.7	5.3	.75
EDI-3 Scales						
Drive for Thinness (7)	9.5	7.7	.91	22.5	8.6	.91
Bulimia (8)	4.8	5.7	.87	17.8	7.4	.89
Body Dissatisfaction (10)	18.2	10.9	.92	37.4	11.8	.92
Low Self-Esteem (6)	4.1	4.3	.84	15.0	5.1	.86
Personal Alienation (7)	4.8	4.4	.78	17.4	5.4	.81
Interpersonal Insecurity (7)	5.0	4.1	.74	18.0	5.1	.78
Interpersonal Alienation (7)	5.4	4.0	.71	17.8	4.9	.74
Interoceptive Deficits (9)	6.5	5.9	.85	23.0	7.1	.86
Emotional Dysregulation (8)	4.3	3.9	.69	17.4	5.1	.73
Perfectionism (6)	11.2	5.0	.73	22.6	5.6	.75
Asceticism (7)	5.5	3.9	.65	17.6	5.0	.68
Maturity Fears (8)	8.2	5.1	.79	23.2	6.1	.81
EDI-3 Composites						
Eating Disorders Risk	32.7	21.4	.91	77.7	24.7	.96
Ineffectiveness	9.0	8.1	.89	32.4	10.0	.91
Interpersonal Problems	10.5	7.0	.82	35.8	8.7	.84
Affective Problems	10.8	8.5	.86	40.4	10.6	.88
Overcontrol	16.7	7.3	.77	40.2	8.6	.79
General Maladjustment	55.2	27.7	.94	172.0	34.5	.95

Note. Weighted item scores for the EDI-2 and EDI-3 range from 0-3 and 0-4, respectively; for both measures, unweighted item scores range from 1-6 (i.e., the original response scale). Alpha coefficients for the EDI-3 composites were estimated using a formula recommended by Nunnally and Bernstein (1994, p. 268). Values in parentheses indicate the number of items on each scale.

identifying "weight preoccupied" females (Garner, 1991), comparable to the 10% reported by Garner for a non-clinical college sample. Also, 21% had EDI-3 Drive for Thinness scores at or above the EDI-3 cutoff, similar to the 22% in Garner's (2004) non-clinical college sample.

Consistent with findings reported previously for female college students (Van Strien & Ouwens, 2003), internal consistency reliabilities for EDI-2 scales derived from unweighted item scores were generally higher than their weighted counterparts (median alphas = .81 and .76, respectively). The differences in reliabilities were slightly lower for EDI-3 scales computed from the unweighted and weighted item scores (median alphas = .81 and .78, respectively). This was anticipated; as noted earlier, EDI-3 weighted scoring compresses the range for each item from six to five points, whereas EDI-2 weighted scoring truncates the range to four points.

Table 2
Varimax Component Loadings and Interpersonal Statistics for the IASR-B5 Octant Scales

	Loadings		Vector	Angular	Cosine
Octant Scale	DOM	LOV	Length	Placement (°)	Difference[1]
PA (90°)	.81	.10	.82	106	0.96
BC (135°)	.67	-.35	.75	140	1.00
DE (180°)	.35	-.76	.84	178	1.00
FG (225°)	-.32	-.82	.88	224	1.00
HI (270°)	-.78	-.37	.87	267	1.00
JK (315°)	-.76	.02	.76	295	0.94
LM (360°/0°)	-.47	.61	.77	345	0.97
NO (45°)	.24	.79	.82	40	1.00
% of Variance	34.9	32.3			

PA = Assured-Dominant; BC = Arrogant-Calculating; DE = Cold-hearted; FG = Aloof-Introverted; HI = Unassured-Submissive; JK = Unassuming-Ingenuous; LM = Warm-Agreeable; NO = Gregarious-Extraverted. Eigenvalues = 2.8, 2.5, 0.8, 0.6, 0.5, 0.3, 0.2, and 0.2.

[1]Correlation between an octant scale's predicted and empirical angular placements after rotating the components to maximum convergence with the theoretical coordinates.

Circumplex Analyses

Analyses of the IASR-B5, summarized in Table 2, indicate that the pattern of correlations among the octant scales was consistent with geometric assumptions for circularity (Gurtman & Pincus, 2003). Specifically: a) application of the parallel analysis criterion (Thompson, 2004) to the eigenvalue distribution from a principal components analysis (see bottom of Table 2) was supportive of the expected two-component solution; b) the two components were uncorrelated (r = -.02; oblimin rotation) and c) similar in size; d) the octant scales' vector lengths were comparable; and e) their angular placements—following rotation to maximum convergence (least-squares criterion)—were generally model-consistent, as indicated by the large cosine differences (mean = .98).

Interpersonal statistics for the EDI-2 and EDI-3 measures are displayed in Table 3. Scales with VLs ≥ .30, a heuristic threshold indicating that a variable has appreciable interpersonal content (Gurtman, 1991), are highlighted in the table and, for the EDI-3, plotted in Figure 2.

Table 3
Interpersonal Statistics for the EDI-2 Scales, EDI-3 Scales, and EDI-3 Composites

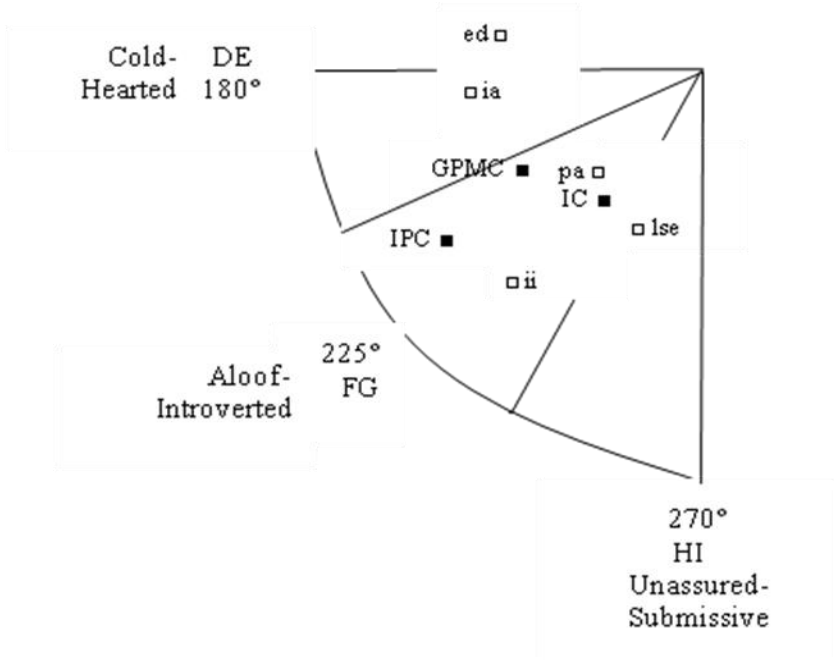
	Weighted Scores		Unweighted Scores	
Measure	VL	AP	VL	AP
EDI-2 Scales				
Drive for Thinness	.08	303.0	.06	293.8
Bulimia	.04	20.8	.02	265.4
Body Dissatisfaction	.11	298.4	.13	294.1
Ineffectiveness	.20	241.7	**.32**	245.4
Perfectionism	.19	101.4	.19	103.4
Interpersonal Distrust	**.36**	202.7	**.44**	215.6
Interoceptive Awareness	.11	252.5	.16	252.8
Maturity Fears	.15	237.2	.21	240.4
Asceticism	.13	215.3	.19	220.0
Impulse Regulation	.28	167.0	**.30**	180.0
Social Insecurity	**.44**	223.9	**.53**	226.8
EDI-3 Scales				
Drive for Thinness	.07	295.4	.06	293.8
Bulimia	.05	348.3	.03	288.4
Body Dissatisfaction	.12	290.7	.12	292.5
Low Self-Esteem	.29	252.6	**.33**	250.7
Personal Alienation	.25	233.1	**.30**	232.7
Interpersonal Insecurity	**.54**	231.1	**.59**	234.6
Interpersonal Alienation	**.38**	187.8	**.40**	190.2
Interoceptive Deficits	.14	251.4	.20	182.6
Emotional Dysregulation	**.32**	165.0	**.32**	172.0
Perfectionism	.20	103.2	.19	103.4
Asceticism	.16	213.4	.18	209.5
Maturity Fears	.18	244.4	.21	240.4
EDI-3 Composites				
Eating Disorders Risk	.10	299.5	.09	291.9
Ineffectiveness	.28	243.4	**.32**	242.0
Interpersonal Problems	**.50**	213.8	**.53**	217.3
Affective Problems	.18	198.1	.22	206.7
Overcontrol	.13	140.2	.14	149.6
General Maladjustment	.29	215.5	**.33**	218.8

Note. VL (Vector Length) = $(DOM^2 + LOV^2)^{½}$ and AP (Angular Placement) = arctangent (DOM/LOV), where DOM and LOV are correlations between scores on the EDI "target" variables and the superordinate interpersonal circumplex scales. VLs ≥ .30 are shown in bold type.

As predicted, none of the VLs for the eating disorder risk scales—Drive for Thinness, Bulimia, Body Dissatisfaction—were ≥ .30, and the highest VLs were for those scales and composites dealing specifically with interpersonal concerns and problems (e.g.,

interpersonal/social insecurity, interpersonal distrust and alienation). APs for the eight variables with VLs ≥ .30 ranged from 172° (DE; Cold-Hearted) to 251° (HI; Unassured-Submissive), with an interpersonal "center" in octant FG (Aloof-Introverted). Within this 79° range, however, there was some variation in angular placement. One EDI interpersonal "mini-cluster," centered on octant scale DE (180°), suggests that interpersonal coldness and hostility are associated with poor affective and behavioral control, particularly in interpersonal situations, and an inability or unwillingness to trust and reach out to others. A second mini-cluster, straddling octant scales FG (225°) and HI (270°), represents the low-self esteem, social insecurity, and withdrawal that co-occur frequently with eating disorder symptoms. Finally, the EDI-3 composites were located between and roughly equidistant from the two mini-clusters.

Figure 2
Interpersonal circumplex projections for EDI-3 scales and composites with vector lengths ≥ .30.



Note. Black squares (■) are EDI-3 composites: GPMC = General Psychological Maladjustment; IC = Ineffectiveness; IPC = Interpersonal Problems. White squares (□) are scales: ed = Emotional Dysregulation; ia = Interpersonal Alienation; pa = Personal Alienation; ii = Interpersonal Insecurity; lse = Low Self-Esteem.

Multiple Regression Analyses

Prior to regressing the 18 EDI-3 measures (12 scales + 6 composites) on the IASR-B5 five-factor scales, we examined the matrix of zero-order correlations. Neuroticism was

correlated significantly ($p < .05$) with all 18 measures, Conscientiousness with 11, DOM and LOV with nine each, and Openness to Experience with two. The correlations of Neuroticism with the EDI-3 measures were all positive, whereas all significant correlations involving the other five-factor scales (except with EDI-3 Perfectionism) were negative. Finally, examination of univariate statistics, bivariate correlations, and standardized residual scatterplots confirmed that the EDI-3 and IASR-B5 data met the assumptions for multiple linear regression (i.e., normally distributed, linearly associated, and homoscedastic).

Results for the simultaneous multiple regression analyses are displayed in Table 4. (Openness to Experience made no significant contributions to the EDI-3 scales and composites and is not included in the table.) As expected, the overall proportions of variance accounted for by the IASR-B5 five-factor scales (i.e., the R^2s) were lower for the three eating disorder symptom scales and the Eating Disorders Risk Composite than for the general psychological maladjustment measures. Neuroticism made statistically significant positive contributions to all of the EDI-3 scales and composites and, with one exception, was the only scale that explained significant variance in the three eating disorder symptom scales.

Table 4
Standardized Partial Regression Coefficients and R^2s for Simultaneous Multiple Regressions of the EDI-3 Scales and Composites on the IASR-B5 Domain Scales

	IASR-B5 Domain Scales								
Criterion Variables	DOM		LOV		C		N		R^2
EDI-3 Scales									
Drive for Thinness							.399	***	.161
Bulimia					-.133	*	.330	***	.123
Body Dissatisfaction							.306	***	.115
Low Self-Esteem	-.217	***			-.224	***	.468	***	.365
Personal Alienation	-.145	*			-.229	***	.435	***	.319
Interpersonal Insecurity	-.481	***	-.333	***			.160	**	.394
Interpersonal Alienation			-.358	***			.213	**	.219
Interoceptive Deficits					-.146	*	.433	***	.236
Emotional			-.233	***	-.137	**	.377	***	.285
Dysregulation									
Perfectionism			.161	*	.122	*	.384	***	.199
Asceticism					-.176	**	.385	***	.202
Maturity Fears					-.224	***	.284	***	.171
EDI-3 Composites									
Eating Disorders Risk			.125	*			.386	***	.160
Ineffectiveness	-.190	**			-.239	***	.476	***	.377
Interpersonal Problems	-.340	***	-.396	***			.214	***	.343
Affective Problems					-.176	**	.467	***	.292
Overcontrol							.473	***	.235
General Maladjustment	-.148	*	-.155	**	-.168	**	.503	***	.387

Note. DOM = Dominance; LOV = Love; C = Conscientiousness; N = Neuroticism. Regression coefficients are shown only for statistically significant predictors ($p < .05$). Openness to Experience is omitted because it made no significant unique contributions to the EDI-3 scales and composites. * $p < .05$; ** $p < .01$; *** $p < .001$.

Conscientiousness made statistically significant contributions—all negative except for Perfectionism—to seven of the nine psychological maladjustment scales and three composites. As anticipated, DOM and LOV explained the most unique variance in the EDI-3 scales and composites specific to interpersonal problems, and the generally negative regression coefficients for DOM and LOV are consistent with angular placements of the EDI-3 measures in the lower left quadrant (180° to 270°) of the interpersonal circumplex (Figure 2).

DISCUSSION

The results of this study illustrate the utility of a combined interpersonal circumplex/five-factor model perspective on the psychometric properties of the EDI-3. The circumplex analyses provided insight into the interpersonal substrate of scales and composites designed to assess broader psychopathological tendencies associated with disordered eating, whereas the five-factor analyses highlighted the contributions of Neuroticism and Conscientiousness to the EDI-3 scales and composites.

The substantial contributions of Neuroticism—across the EDI-3 scales and composites—were expected. Brookings and Wilson (1994) reported similar results for the original EDI (Garner & Olmsted, 1984), and current models of eating disorders (e.g., Tylka & Subich, 2004) emphasize the role of "negative affect" (i.e., neuroticism and low self-esteem) in disordered eating. Conscientiousness explained less unique variance than did Neuroticism, but did make significant (negative) contributions to general psychopathology scales dealing with primarily intrapersonal concerns (e.g., Personal Alienation, Maturity Fears). Finally, and as predicted, DOM and LOV were the largest FFM contributors to the interpersonally-oriented scales (Interpersonal Insecurity, Interpersonal Alienation) and composite (Interpersonal Problems).

In the circumplex analyses, the angular placements of scales with substantial interpersonal loadings were less variable than anticipated. The interpersonal center for these measures was octant FG (225°), but there was evidence as well for two smaller mini-clusters of scales: one reflecting hostility and poor impulse regulation, the other related to low self-esteem, insecurity, and withdrawal from others (see also Madison, 1997). Affect and behavior associated with these interpersonal styles—including depression, loneliness, and social avoidance (Wiggins & Broughton, 1991)—are in turn associated with interpersonal stressors, which among women with bulimia often precede and may trigger binge eating (Steiger, Gauvin, Jabalpurwala, Seguin, & Stotland, 1999). In summary, a number of the EDI-3 measures covary with *interpersonal deficits* (e.g., social isolation, inadequate or impoverished interpersonal relationships) that are central to interpersonal vulnerability models of eating disorders (Wilfley et al., 2005).

As predicted, neither the EDI-3 symptom scales nor the symptom composite had appreciable loadings in the interpersonal circumplex space. One implication of these findings is that interpersonal variables may be useful primarily for predicting satisfaction with and continuation of treatment, rather than direct assessment of eating disorder symptoms per se. This is consistent with recent research suggesting that a) the relationship between interpersonal variables and psychopathology is complex and variable, even within diagnostic categories (Hopwood, Clarke, & Perez, 2007); and b) "...a targeted interpersonal approach on the part of the therapist has the potential to enhance the alliance and limit treatment dropout" (Ambwani & Hopwood, 2009, p. 248).

The principal limitation of this study is its cross-sectional design, which precluded testing hypotheses on the causal priority of interpersonal problems and disordered eating. One suggestion for future research, then, is longitudinal assessment of interpersonal characteristics and eating disorder symptoms. Such studies would enable identification of interpersonal "markers" for subsequent eating disorders, and inform efforts to reduce their incidence. A second limitation of this study is that, because the participants were non-patients, it was not possible to explore possible differences in circumplex projections among the various eating disorder syndromes and subtypes in the DSM-IV (American Psychiatric Association, 1994). Another direction for future research, then, is systematic investigation of the link between interpersonal tendencies and diagnosed eating disorders, starting perhaps with the interpersonal mini-clusters suggested by this study.

Finally, we concur with those investigators (e.g., Schoemaker et al., 1994; Van Strien & Ouwens, 2003) who conclude that collapsing or "weighting" EDI item scores may compromise unnecessarily the psychometric properties of the scales, particularly in non-clinical populations. Specifically, scales produced by summing weighted item scores had less variability and lower internal consistency reliabilities than did their unweighted counterparts, and in the circumplex analyses, had lower interpersonal loadings as well. To the extent that the interpersonal content in these scales may function as subtle indicators of incipient but sub-threshold eating problems in college females, for whom problematic eating behavior is quite common (Kurth et al., 1995; Mintz & Betz, 1988), restricting the range of item scores seems counterproductive. Accordingly, we recommend derivation of EDI-3 norms based on unweighted item scores as a supplement to the weighted-score norms reported in the EDI-3 manual (Garner, 2004).

In summary, these results confirm that a) negative affect (i.e., Neuroticism) is the largest overall contributor, among five-factor model traits, to the EDI-3 scales and composites; but b) there is a potentially important role as well for interpersonal variables in the assessment of eating disorders. Specifically, the prominent interpersonal features of the EDI-3 scales identified here—interpersonal alienation, insecurity about approaching and confiding in others, and a general reluctance to form close relationships—are those which in previous studies have been associated with the etiology of eating-related problems, as well as with client/therapist relationship problems leading to dissatisfaction with treatment and higher likelihood of dropout.

REFERENCES

Ambwani, S., & Hopwood, C. J. (2009). The utility of considering interpersonal problems in the assessment of bulimic features. *Eating Behaviors*, *10*, 247-253.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Benjamin, L. S. (1996a). *Interpersonal diagnosis and treatment of personality disorders* (2nd ed.). New York: Guilford.

Benjamin, L. S. (1996b). A clinician-friendly version of the interpersonal circumplex: Structural Analysis of Social Behavior (SASB). *Journal of Personality Assessment*, *66*, 248-266.

Brookings, J. B., & Wilson, J. F. (1994). Personality and family-environment predictors of self-reported eating attitudes and behaviors. *Journal of Personality Assessment*, *63*, 313-326.

Garner, D. M., & Olmsted (1984). *Manual for Eating Disorder Inventory*. Odessa, FL: Psychological Assessment Resources, Inc.

Garner, D. M. (1991). *Eating Disorder Inventory-2 professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Garner, D. M. (2004). *Eating Disorder Inventory-3 professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.

Gurtman, M. B. (1991). Evaluating the interpersonalness of personality scales. *Personality and Social Psychology Bulletin*, *17*, 670-677.

Gurtman, M. B. (1992). Construct validity of interpersonal measures: The interpersonal circumplex as a nomological net. *Journal of Personality and Social Psychology*, *63*, 105-118.

Gurtman, M. G. (1994). The circumplex as a tool for studying normal and abnormal personality: A methodological primer. In R. Plutchik & H. R. Conte (Eds.), *Circumplex models of personality and emotions* (pp. 81-102). Washington, DC: American Psychological Association.

Gurtman, M. B., & Pincus, A. L. (2003). The circumplex model: Methods and research applications. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology (Vol. 2): Research methods in psychology* (pp. 407-428). New York: Wiley.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258-348). Glencoe, IL: Free Press.

Hopwood, C. J., Clarke, A. N., & Perez, M. (2007). Pathoplasticity of bulimic features and interpersonal problems. *International Journal of Eating Disorders*, *40*, 652-658.

Horowitz, L. M. (2004). *Interpersonal foundations of psychopathology*. Washington, DC: American Psychological Association.

Horowitz, L. M., Alden, L. E., Wiggins, J. S., & Pincus, A. L. (2000). *Manual for the Inventory of Interpersonal Problems*. San Antonio, TX: The Psychological Corporation.

Kurth, C. L., Krahn, D. D., Nairn, K., & Drewnowski, A. (1995). The severity of dieting and bingeing behaviors in college women: Interview validation of survey data. *Journal of Psychiatric Research*, *29*, 211-225.

Leary, T. (1957). *Interpersonal diagnosis of personality: A functional theory and methodology for personality evaluation*. New York: Ronald Press.

Madison, J. K. (1997). Interpersonal assessment and therapy of eating disorders: A clinical application of a circumplex model. In R. Plutchik & H. R. Conte (Eds.), *Circumplex models of personality and emotions* (pp. 431-446). Washington, DC: American Psychological Association.

McCrae, R. R., & Costa, P. T., Jr. (1989). The structure of interpersonal traits: Wiggins' circumplex and the five-factor model. *Journal of Personality and Social Psychology*, *56*, 586-595.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*, 175-215.

Mintz, L. B., & Betz, N. E. (1988). Prevalence and correlates of eating disordered behaviors among undergraduate women. *Journal of Counseling Psychology*, *35*, 463-471.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Nylander, I. (1971). The feeling of being fat and dieting in a school population: An epidemiologic interview investigation. *Acta Socio-Medica Scandinavica*, *3*, 17-26.

Schoemaker, C, Van Strien, T., & Van der Staak, C. (1993). Validation of the Eating Disorders Inventory in a nonclinical population using transformed and untransformed responses. *International Journal of Eating Disorders*, *15*, 387-393.

Steiger, H., Gauvin, L., Jabalpurwala, S., Seguin, J. R., & Stotland, S. (1999). Hypersensitivity to social interactions in bulimic syndromes: Relationship to binge eating. *Journal of Consulting and Clinical Psychology*, *67*, 765-775.

Theander, S. (2004). Trends in the literature on eating disorders over 36 years (1965-2000): Terminology, interpretation, and treatment. *European Eating Disorders Review*, *12*, 4-17.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Tracey, T. J. G. (2000). Analysis of circumplex models. In H. E. A. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 641-664). San Diego: Academic Press.

Trapnell, P. D., & Campbell, J. D. (1999). Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology*, *76*, 284-304.

Trapnell, P. D., & Wiggins, J. S. (1990). Extension of the Interpersonal Adjective Scales to include the big five dimensions of personality. *Journal of Personality and Social Psychology*, *59*, 781-790.

Tylka, T. L., & Subich, L. M. (2004). Examining a multidimensional model of eating disorder symptomatology among college women. *Journal of Counseling Psychology*, *51*, 314-328.

Van Strien, T., & Ouwens, M. (2003). Validation of the Dutch EDI-2 in one clinical and two nonclinical populations. *European Journal of Psychological Assessment*, *19*, 66-84.

Wiggins, J. S. (1995). *Interpersonal Adjective Scales professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Wiggins, J. S., & Broughton, R. (1985). The interpersonal circle: A structural model for the integration of personality research. In R. Hogan & W. Jones (Eds.), *Perspectives in personality* (Vol. 1, pp. 1-47). Greenwich, CT: JAI Press.

Wiggins, J. S., & Trapnell, P. D. (1996). A dyadic-interactional perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 88-162). New York: Guilford.

Wiggins, J. S., & Trobst, K. K. (2002). The Interpersonal Adjective Scales: Big Five version (IASR-B5). In B. DeRaad & M. Perugini (Eds.), *Big five assessment* (pp. 263-280). Seattle: Hogrefe and Huber.

Wilfley, D., Stein, R., & Welch, R. (2005). Interpersonal psychotherapy. In J. Treasure, U. Schmidt, & E. van Furth (Eds.), *The essential handbook of eating disorders* (pp. 137-154). West Sussex, England: Wiley.