A Bifactor and Item Response Theory Analysis of the Eating Disorder Inventory-3

Jeffrey B. Brookings#, Dennis L. Jackson & David M. Garner

Journal of Psychopathology and Behavioral Assessment

ISSN 0882-2689

J Psychopathol Behav Assess
DOI 10.1007/s10862-020-09827-2

Springer

Journal of Psychopathology and Behavioral Assessment
https://doi.org/10.1007/s10862-020-09827-2

A Bifactor and Item Response Theory Analysis of the Eating Disorder Inventory-3

Jeffrey B. Brookings[1] · Dennis L. Jackson[2] · David M. Garner[3]

Abstract
The Eating Disorder Inventory-3 (EDI-3; Garner, 2004) is a 91-item, self-report measure scored on 12 scales (three Eating Disorder Risk scales, nine Psychological scales) and six composites. A sample of 1206 female eating disorder patients was divided randomly into calibration ($n = 607$) and cross-validation ($n = 599$) samples for confirmatory factor analyses. A bifactor model best fit the data in both samples, but a model with second-order factors corresponding to the risk and psychological scales approached the fit of the bifactor model. Rasch analysis identified only two items whose level of misfit showed a lack of coherence with other scale items (the only items referring to drug and alcohol use), there were no items with reversed or "disordered" response categories, and only five items had sub-threshold estimated discrimination values. Overall, the results were supportive of the EDI-3's psychometric properties and consistent with interpretive guidelines presented in the test manual.

Keywords Eating disorders · EDI-3 · Confirmatory factor analysis · Rasch analysis

The Eating Disorder Inventory (EDI; Garner, Olmsted, & Polivy, 1983) was designed primarily as a clinical instrument that produces psychological profiles useful for case conceptualization and treatment planning for those with confirmed or suspected eating disorder diagnoses. However, the EDI has never been intended as a diagnostic instrument; within each diagnostic group, there is extraordinary variability on EDI psychological scales consistent with psychological traits that are theoretically unrelated or only weakly related to diagnostic markers such as behavioral symptoms and body weight. The measure has also been used as a research tool for assessing areas of psychopathology of interest in theory-testing, identifying meaningful patient subgroups, and assessing treatment outcome.

Over time, the EDI has gained popularity among eating disorder clinicians and researchers and has been revised twice. The original EDI was comprised of 64 self-report items responded to on a six-point scale ranging from "Never" to "Always." Item scores were then collapsed to a four-point scale (0–3) and the items assigned to eight scales which assessed individual differences in eating disorder risk (three scales) and associated psychological features (five scales). The "eating disorder risk" scales tap eating-disorder-specific symptoms and were later combined into a "composite" used to screen non-clinical groups to determine the presence of attitudes or behaviors that may indicate risk of clinical or subclinical eating disorders. The five psychological scales assess traits that are highly relevant to, but not specific to, eating disorders. The first revision (EDI-2; Garner, 1991) added 27 items and increased the number of scales to 11. The subsequent EDI-3 revision (Garner, 2004) was guided by the evolution of theoretical models in the field since the original measure was introduced and by evaluating the clinical constructs underlying the original scales. The EDI-3 validation retained the 91 EDI-2 items and examined the relationships among items by applying exploratory factor analysis (EFA) to normative and eating disorder samples. The analysis yielded item clusters largely conforming to the EDI-2 scale structure but changed the scale assignment for some items, added a 12th

✉ Jeffrey B. Brookings
jbrookings@wittenberg.edu

Dennis L. Jackson
djackson@uwindsor.ca

David M. Garner
dm.garner@gmail.com

[1] Psychology Department, Wittenberg University, Springfield, OH 45504, USA

[2] Department of Psychology , The University of Windsor , 401 Sunset Ave, N9B 3P4 Windsor, Canada

[3] River Centre Foundation, 5019 Rolandale Ave, Toledo, OH 43623, USA

Published online: 28 July 2020

Table 1 *Descriptive Statistics for the EDI-3 Scales and Composites (N = 1206)*

Scale/Composite	M	SD	Skew	α Total[1]	PHP[2]	RES[3]
Risk Scales						
Drive for Thinness	20.44	7.41	−1.17	.89	.87	.91
Bulimia	12.97	8.95	.34	.92	.92	.92
Body Dissatisfaction	27.62	10.34	−.66	.90	.90	.91
Psychological Scales						
Low Self-Esteem	13.52	6.02	−.26	.86	.86	.86
Personal Alienation	14.69	6.42	−.13	.82	.81	.85
Interpersonal Insecurity	12.33	6.03	.04	.81	.80	.83
Interpersonal Alienation	11.06	5.55	.18	.76	.76	.78
Interoceptive Deficits	17.68	8.90	−.04	.88	.88	.88
Emotional Dysregulation	8.97	6.29	.70	.77	.76	.80
Perfectionism	13.31	6.00	−.16	.80	.80	.81
Asceticism	12.52	6.33	.07	.77	.76	.78
Maturity Fears	11.72	7.05	.64	.83	.85	.80
Composites						
Eating Disorder Risk	61.04	21.81	−.65	.93	.92	.94
Ineffectiveness	28.21	11.75	−.26	.91	.90	.92
Interpersonal Problems	23.39	10.17	.06	.85	.84	.86
Affective Problems	26.64	13.53	.15	.89	.87	.90
Overcontrol	25.83	10.52	.01	.83	.82	.84
General Psychological Maladjustment	115.78	41.08	−.10	.95	.96	.97

Note. Scale and composite statistics are based on item sums. For all scales, higher scores reflect greater distress

[1] Total sample ($N = 1206$)

[2] Partial Hospitalization Program sample ($n = 821$)

[3] Adolescent Residential Program sample ($n = 385$)

scale, and expanded the item scores from four to five points (now 0–4). In addition to the 12 primary scales, six composite scales and three response style indicators were added (see Table 1 for a list of EDI-3 scales and composites).

Like its predecessors, the EDI-3 has received generally favorable reviews. One reviewer (e.g., Cumella, 2006) praises its "…superior section on test interpretation" (p. 117) but notes that questions remain about the EDI-3 factor structure. This is an important issue: The rich interpretive material in the EDI-3 manual has clinical utility only to the extent that the test's purported factor structure is confirmed. Garner (2004) reported the results of separate exploratory factor analyses of the eating disorder risk and psychological scale items and a confirmatory factor analysis (CFA) of the nine psychological scales but did not conduct item level CFAs. A recent item CFA of the EDI-3 for adult Danish patients and non-patient controls was generally supportive of its purported factor structure. Specifically, the best-fitting models were: a) a first-order model with correlated factors corresponding to the 12 scales; and b) a second-order model with two global factors: Eating Disorder Risk and Psychological Disturbance (Clausen, Rosenvinge, Friborg, & Rokkedal, 2011).

Clausen et al.'s (2011) study was the first item CFA of the EDI-3 and the first to assess a broad set of first- and second-order models (though, for reasons not stated, they omitted the six composites from their analyses). However, because their findings derived from a translated version of the EDI-3 and their data were exclusively from Danish participants, the primary objective of this study was to complete the first item CFA of the EDI-3 in a sample of English-speaking eating disorder patients, starting with Clausen et al.'s best-fitting first- and second-order factor models as initial targets. Furthermore, the positive and generally large correlations among the EDI-3 scales and composites (e.g., Garner, 2004) suggest the influence of a general psychological distress factor. If so, a bifactor model (Reise, 2012) might provide a clearer picture of the EDI-3 s dimensionality. In a comprehensive comparison of bifactor and second-order factor models, Chen, West, and Sousa (2006) note that one advantage of the former is that it allows the researcher to more directly assess

variance in each first-order factor—and the items that load on them—that is independent of the general distress factor. What this implies for the current study is that support for Garner's (2004) recommendation to begin EDI-3 clinical interpretation with the 12 primary scales obtains if the first-order factors—corresponding to the 12 scales—retain substantial variance after controlling for the influence of the bifactor. Therefore, a second objective of this research was to evaluate the fit of a bifactor model, relative to first-order models and the second-order models fitted by Clausen et al. (2011), and to assess directly the loadings of EDI-3 items on the bifactor and their respective first-order "content" factors.

Clausen et al. (2010) reported that, based upon inspection of modification indices, model misfit was attributable mainly to cross-loading items and item pairs with large correlated errors, but they reported no details. Accordingly, the third objective of this research was an evaluation of item-level sources of model misfit, guided by substantive and statistical criteria. And to supplement the information provided by the CFAs, we ran item response theory (IRT) analyses using the one-parameter Rasch (1960) rating scale model. IRT analysis provides detailed information about the strengths and weaknesses of scales and items, including item fit to the model, estimated item discrimination values, and whether the response categories function as intended.

The rationale for the current study using data from a heterogenous clinical sample of eating disorder patients is to determine if the psychometric properties of the EDI-3 are consistent with the interpretive guidelines published in the manual and the findings reported by Clausen et al. (2011). Results bearing on the factor structure and item characteristics of the EDI-3 in a clinical sample has both clinical and research utility since it adds to the body of literature on the measure's validity. In sum, the overarching aim of the current study is to determine if CFA and IRT analyses provide evidence for the construct validity of the 12 scales as originally configured in the EDI-3 manual (2004) and supported by Clausen, et al.'s (2011) results.

Method

Participants

The initial sample comprised 1317 consecutive admissions between October 2005 and December 2014 to a specialized eating disorder treatment facility in the upper Midwest in the United States. Patients completed the EDI-3 as part of the intake process to either a partial hospitalization (PHP; n = 868) or adolescent residential treatment program (RES; n = 449). Approximately 7% of the patients admitted to the PHP program were under the age of 18 years old. The average ages for those admitted to the PHP and RES were 25.7 years and 15.7 respectively. For those electing to specify race, the sample was predominantly Caucasian (94%) with a small minority of Hispanic (3%), Asian (2%) and Native American (1%). Although SES and educational data were not available on the sample, patients served by the treatment facility were largely local/regional representing indigenous working-class and professional families with good insurance plans including Medicaid. Because there were only 59 male patients, their data were excluded from the subsequent analyses. For the remaining 1258 female patients, we examined distributions of missing item responses and removed 52 cases with six or more non-answered items. T-tests comparing the retained and deleted cases (n = 1206 and 52, respectively) showed that those deleted were significantly younger ($p < .05$): M = 20.0 (SD = 8.4) vs 22.6 (SD = 8.9), but there were no differences in pre- and post-treatment weight change, BMI, or bingeing and vomiting frequency, and the contingency between retention/exclusion and DSM-5 diagnosis was non-significant (χ^2 = 0.032). The final sample of 1206 consisted of 821 PHP mostly adult patients (54 or 6.6% were under the age of 18 years) and 385 RES patients (100% < 19 years old). The mean age of the final sample was 22.6 years (SD = 8.9); PHP = 25.7 years (SD = 9.1) and RES = 15.8 years (SD = 1.6); range: PHP = 11.4–74.3 years and RES = 11.3–18.2 years.

In the final sample of 1206 cases, missing item responses were imputed for 407 cases with one to five missing responses using Expectation Maximization. Among those 407, 250 (60.4%) had only one missing response and considering the entire data matrix, less than 1% of data were missing. The final sample consisted of the following DSM-5 (American Psychiatric Association, 2013) diagnostic groups: Anorexia Nervosa-Restricting Type (n = 341); Anorexia Nervosa-Bingeing/Purging Type (n = 259); Bulimia Nervosa (n = 404); Other Specified Feeding or Eating Disorder (n = 155); Binge Eating Disorder (n = 42); Avoidant/Restrictive Food Intake Disorder (n = 2); and Unspecified Feeding or Eating Disorder (n = 3). All diagnoses were made according to the DSM-5 by a licensed clinician and reviewed by research staff to ensure that the clinical diagnoses were consistent with the diagnostic criteria.

The focus of the current paper is the examination of the psychometric properties of the EDI-3 in a clinical eating disorder sample and is not intended to provide information specific to diagnostic subgroups; therefore, all 1206 cases were randomly assigned to one of two subsamples to cross-validate the best-fitting factor models.

The retrospective chart review was approved by the Clinic Institutional Review Board in compliance with Health Insurance Portability and Accountability Act guidelines. Use of the data conformed to HIPAA standards for use of de-identified, archival data. Patient names and identifiers were removed prior to conducting all analyses. Informed consent to participate in archival research was obtained from all individual participants included in the study.

Measure

Test takers respond to the 91 EDI-3 items on a six-point scale ranging from "Never" (0) to "Always" (5). Sixty-six items are negatively phrased (e.g., "I feel extremely guilty after overeating") and 25 items are positively phrased (e.g., "I eat sweets and carbohydrates without feeling nervous"). For all items, the two least symptomatic options are assigned a score of 0, with scores of 1, 2, 3, and 4 given to progressively more symptomatic responses. The rationale for using the 0–4 scoring system, rather than a 1–6 scoring system, is both rational-theoretical and empirical, derived from the assumption that EDI item scaling is continuous only for responses weighted 1 to 4 (Garner, 2004). With a 1–6 scoring system, it is possible for three responses in the non-symptomatic direction to receive the same empirical weight as one extreme response in the symptomatic direction. The EDI-3 manual suggests that the two responses in the non-symptomatic direction should not contribute to the total scale score reflecting psychopathology because it is not intuitive or rational for a respondent to receive different scores for "rarely" or "never" in the non-symptomatic direction (Garner, 1991). The EDI was developed to provide a psychological profile for clinical samples. The lower reliabilities reported for nonclinical samples are expected; most item distributions are negatively skewed because they are less relevant to most individuals in nonclinical samples. Nevertheless, the frequency with which the EDI has been used to address theoretical questions in nonclinical groups has led some to adopt the 1–6 scoring system (e.g., Keel, Baxter, Heatherton & Joiner, 2007). The 0–4 scoring system was selected for the EDI-3 because it retains the heuristic of the original scoring format but expands the range of scores, which improves the psychometric qualities of the EDI-3 primarily for nonclinical samples. After reverse-scoring the positively phrased items, the 90 truncated item scores (item #71 is not scored) are summed to produce 12 raw scale scores, which are then converted to T-scores using diagnostic group norms in the EDI-3 manual (Garner, 2004). Finally, five composite T-scores are formed by summing combinations of scale scores. For example, the Interpersonal Problems (IPC) composite is derived by combining Interpersonal Insecurity (II) and Interpersonal Alienation (IA) and the General Psychological Maladjustment Composite is formed by combining all eight psychological content scales. Descriptive statistics for the Eating Disorder Risk, Psychological and Composite scales are reported in Table 1.

Confirmatory Factor Analyses (CFA)

The 1206 cases were divided randomly into calibration (n = 607) and cross-validation (n = 599) samples. The first sample was used to fit all a priori models and models with post-hoc modifications, which were based on examining modification indices for correlated residuals and considering the plausibility of each modification. Models tested on sample 1 were cross-validated on sample 2 data. Subsequently, the best fitting models were fit to the entire sample to provide optimal parameter estimates.

All CFAs were performed with EQS. 6.1 (Bentler & Wu, 2002). From previous research on the EDI-3, we derived and evaluated a targeted set of factor models. Comparisons of model fit were based upon the following measures: Chi-square; Comparative Fit Index (CFI); Tucker-Lewis Index (TLI); Root Mean Square Error of Approximation (RMSEA), and Akaike's Information Criterion (AIC). Due to multivariate kurtosis (Normalized Mardia estimate = 148.1), the Satorra-Bentler corrected χ^2 and incremental fit indices using this correction were also reported. Adequate fit is indicated by values $\geq$ .95 for the CFI and TLI (Hu & Bentler, 1999) and $\leq$ .06 for the RMSEA (Browne & Cudeck, 1993). The AIC is a model comparison index that considers both model fit and model parsimony (Brown, 2015). Smaller values indicate better fit. We anticipated lower fit measures based on the size of the model relative to the sample size when our sample was split for cross-validation purposes. Jackson, Voth and Frey (2013) recommended minimum sample size requirements for larger models under ideal conditions (namely multivariate normality and that the true model has been identified). They found for models with 12 latent variables a sample size between 400 and 1000 would be necessary, depending upon loading size. Our average factor loading (.64) is between their two conditions of .80 (minimum N = 400) and .40 (minimum N = 1000). We chose to rely more heavily on the RMSEA values since they are less biased by sample size than incremental measures, especially for large samples (Jackson, 2003; Rigdon, 1996).

Base Factor Models

Model 0: Null Model

Model 1: 12 Correlated Factors For this model, which was the best-fitting model in Clausen et al.'s (2011) study and corresponds to Garner's (2004) recommended starting point for clinical use of the EDI-3, we estimated all correlations among the 12 content factors.

Model 2: 12 Correlated Factors with 10 Correlated Errors for Inconsistency Scale Items This model was the same as Model 1, but we allowed 10-pairs of error variances to correlate. The 10 pairs comprised the Inconsistency Index for the EDI-3 (Garner, 2004).

Model 3: Two Second-Order Factors High intercorrelations among the first-order latent variables and Garner's (2004) partitioning of the 12 scales into risk and psychological scales

for clinical interpretation led us to examine a model where the two sets of scales load on separate second-order factors.

Model 4: Correlated Content Factors plus a Bifactor A bifactor models shared variance among the items that is separable from the variance associated with their content factors and in our study could be interpreted as a general distress factor. In this model, all 12 first-order latent variables were allowed to correlate as in Model 1, and we added a bifactor, orthogonal to the 12 content factors, with all 90 scored items loading on it (Fig. 1).

Item Response Theory (IRT) Analyses

For the IRT analysis, we used the one-parameter Rasch rating-scale model (Rasch, 1960), which calibrates item and person levels in logits or log-odds units. The model requires that individual scales are unidimensional and that items are locally independent (i.e., that shared item variance is accounted for by the unidimensional construct). We assessed the fit of each item on the 12 EDI-3 content scales analyses, using Winsteps Version 3.92.1 (Linacre, 2016). We examined the following indicators—applying frequently recommended rules of thumb—to evaluate the fit of the data and persons to the Rasch model.

Infit/Outfit Information about unexpected responses close to (Infit statistics) or distant from (Outfit statistics) estimated person and item levels are reported here as mean squares. Values between 0.5 (redundancy) and 1.5 (noise) are considered acceptable; items with mean squares >2.0 severely degrade the psychometric quality of a scale (Wright & Linacre, 1994).

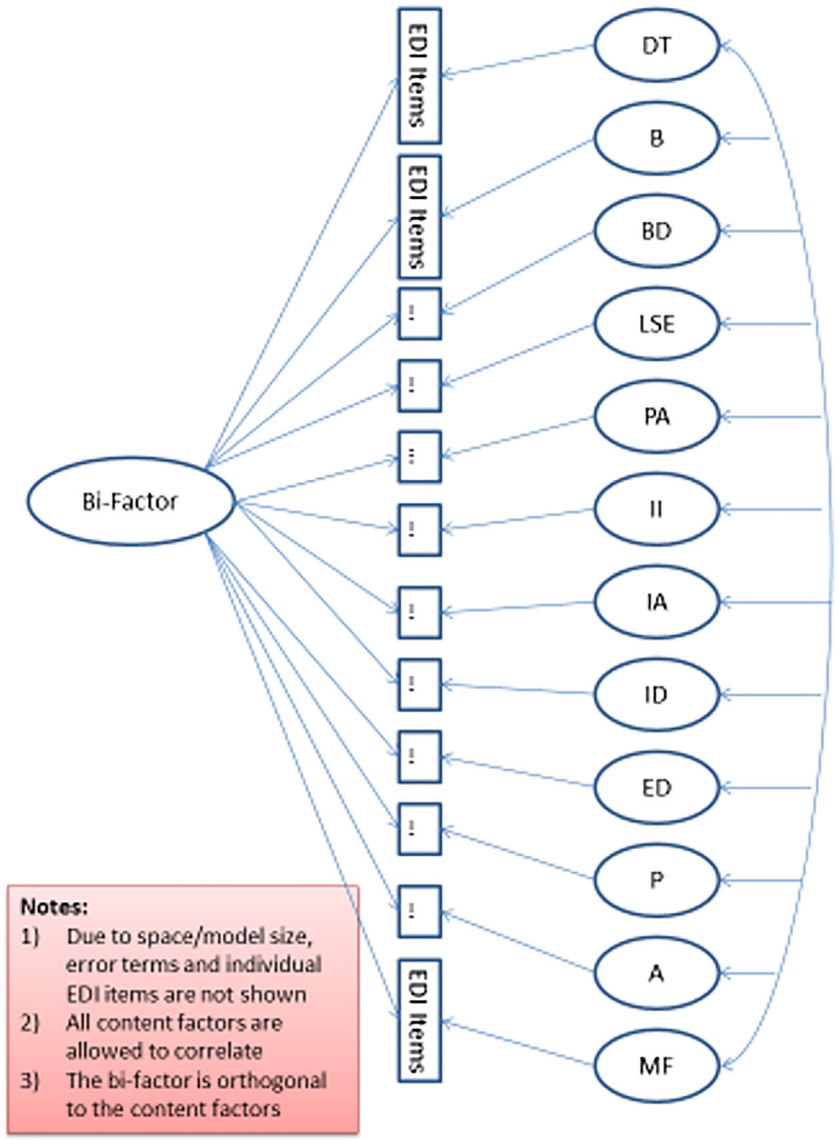

Fig. 1 Bifactor model for the EDI-3

Response Category Thresholds and Ordering We examined the points at which the selection of two adjacent response options was equally probable to verify appropriate category use and that response category scores increased with estimated trait levels (i.e., that there were no "disordered" categories).

Person and Item Separation Person reliability, the Rasch analogue to Cronbach's alpha, assesses how well a scale differentiates persons on the latent trait. Additionally, the person separation index estimates, in standard error units, the spread of persons and items on the scale. Person reliability and separation are influenced primarily by the number of items, the range of scores, and sample-item alignment. Item reliability and separation reflect the extent to which item difficulties are differentiated along the trait continuum. They are most sensitive to sample size and score range. Person reliabilities > .80 and item reliabilities > .90 are considered acceptable. For the separation indexes, person values ≥2.00 and item values ≥3.00 indicate adequate separation (Linacre, 2016).

Estimated Item Discrimination Though Rasch analysis directly emphasizes one parameter, difficulty, the Winsteps program produces an estimate of the discrimination parameter featured in two- and three-parameter IRT models. A value of 1.00 indicates that the item discriminates between high and low scorers as expected by the Rasch model, whereas values greater or less than 1.00 are evidence that the item discriminates more or less, respectively, than predicted. Values ≥ .50 are considered adequate (Linacre, 2016).

Results

Descriptive Statistics

Item means (0 to 4-point scale) for the 12 scales and composites ranged from 1.12 to 2.92, all skews were ≤ 1.17 in absolute value, and internal consistency coefficients ranged from .76 to .92. Scale raw score means and standard deviations as well as scale alphas are reported in Table 1. Median alphas for the scales and composites were .82 and .90, respectively. Gleaves, Pearson, Ambwani, and Morey (2014) recommend reporting reliabilities for adults and adolescents separately since they found a tendency for lower reliabilities for adolescents; however, we found that our adolescent sample had similar or slightly higher scale reliabilities except for Maturity Fears (Table 1). All correlations among the 12 scales and six composites were positive and generally moderate to large in magnitude (Tables 2 and 3, respectively).

CFAs for Samples 1 and 2

Table 4 contains fit information for all models in samples 1 and 2.

Model 1: 12 Correlated Factors Model 1 demonstrated close fit (Browne & Cudeck, 1993), RMSEA = .046. All parameter estimates fell within an acceptable range and the signs were consistent with expectations. All factor covariances were significant and the correlation between two of the latent variables, Personal Alienation and Low Self-Esteem, approached unity (r = .92). The next highest correlation was for latent variables Drive for Thinness and Body Dissatisfaction

Table 2 *EDI-3 Scale Correlations (N = 1206)*

Scale	DT	B	BD	LSE	PA	II	IA	ID	ED	P	A	MF
DT	–											
B	.415	–										
BD	.720	.370	–									
LSE	.539	.396	.616	–								
PA	.455	.423	.507	.784	–							
II	.259	.174	.333	.469	.467	–						
IA	.325	.392	.402	.575	.680	.540	–					
ID	.459	.419	.463	.578	.683	.436	.522	–				
ED	.321	.388	.320	.497	.593	.268	.499	.575	–			
P	.332	.239	.274	.288	.316	.183	.340	.356	.214	–		
A	.547	.457	.517	.554	.571	.306	.495	.605	.475	.456	–	
MF	.184	.173	.148	.283	.350	.190	.209	.314	.276	.103	.198	–

Note. All correlations are statistically significant ($p < .05$). DT = Driver for Thinness; B = Bulimia; BD = Body Dissatisfaction; LSE = Low Self-Esteem; PA = Personal Alienation; II = Interpersonal Insecurity; IA = Interpersonal Alienation; ID = Interoceptive Deficits; ED = Emotional Dysregulation; P = Perfectionism; A = Asceticism; MF = Maturity Fears

Table 3 *EDI-3 Composite Correlations (N = 1206)*

Composite	EDR	I	IP	AP	O
EDR	–				
I	.634	–			
IP	.434	.656	–		
AP	.553	.703	.556	–	
O	.557	.527	.432	.546	–
GPM	.658	.884	.777	.851	.742

Note. All correlations are statistically significant ($p < .05$)
EDR = Eating disorder Risk; I = Ineffectiveness; IP = Interpersonal Problems; AP = Affective Problems; O = Overcontrol; GPM = General Psychological Maladjustment

($r = .79$). A few others fell within the .60 to .69 range and the remainder were smaller. In the cross-validation sample, the fit was similar (RMSEA = .048) and again, all measured variables had significant loadings on their respective latent variables, and a similar pattern was observed with respect to latent variable correlations.

Model 1A: 12 Correlated Factors plus Five Select Correlated Errors Inspection of residuals and modification indices for Model 1 (sample 1) led us to add five pairs of correlated error terms that were substantively justifiable. Examples include items with shared content (e.g., items 72 [drugs] and 81 [alcohol] are the only two on the Emotional Dysregulation scale that refer to substance abuse concerns) and items that are oppositely-valenced versions of similar content (e.g., items 2 and 12 on the Body Dissatisfaction scale describe negative and positive self-assessments, respectively, of the respondent's stomach size). This model fit better than Model 1 with lower RMSEA and AIC. All measured variables loaded significantly on their respective latent variables and all five correlated errors were significant and ranged from $r = .18$ to $r = .57$. All correlated errors were significant in sample 2 as well, ranging from $r = .20$ to $r = .65$. As in sample 1, all factor loadings and correlations were significant with signs in the anticipated direction and the fit was slightly worse but still reasonable. The impact of estimating correlated errors on factor correlations was negligible.

Model 2: 12 Correlated Factors with 10 Correlated Errors for Inconsistency Scale Items Model fit was adequate and superior to Model 1, but not as good as Model 1A. In fact, only four of the ten pairs in sample 1 and two of the ten pairs in sample 2 covaried significantly.

Model 3: Two Second-Order Factors Model 3 had two second-order latent variables, one for the risk latent variables and a second for the psychological latent variables. Clausen et al. (2011) reported good fit for this model. For our data, a) all measured variables loaded significantly on their respective first-order factors; b) all first-order factors loaded significantly on their respective second-order factor; and c) the two second-order factors correlated significantly ($r = .70$). However, this model did not fit as well as Models 1, 1A or 2 (see Table 4).

Table 4 *Goodness-of-Fit Indices for the EDI-3 Factor Models: Samples 1 and 2, Five-Point Scoring*

Model	χ^2	df	CFI	TLI	RMSEA	RMSEA LOW	RMSEA HIGH	AIC
M0: Null	30,496.64 31,352.68	4005						
M1: 12 corr factors	8693.36 9114.83	3849	.817 .807	.810 .800	.046 .048	.044 .047	.047 .049	995.37 1416.83
M1A: 12 corr factors, select correlated errors	8137.97 8399.21	3844	.838 .833	.831 .826	.043 .045	.042 .043	.044 .046	449.97 711.21
M2: 12 corr factors, Inconsistency Scale corr errors	8542.46 8895.06	3839	.822 .815	.815 .807	.045 .047	.044 .046	.046 .048	864.46 1217.06
M3: Two 2nd order factors	9102.83 9570.11	3902	.804 .793	.798 .787	.047 .049	.046 .048	.048 .050	1298.83 1766.11
M3A: Two 2nd order factors, select corr errors	8546.15 8853.54	3897	.825 .819	.820.814	.044 .046	.043 .045	.046 .047	752.15 1059.54
M4: Bifactor, corr content factors	7498.63 7797.98	3759	.859 .852	.850 .843	.041 .042	.039 .041	.042 .044	−19.37 279.98
M4A: Bifactor, corr content factors, select corr errors	6933.29 7066.78	3754	.880 .879	.872 .871	.037 .038	.036 .037	.039 .040	−574.71 −441.22

Note: For each model, Sample 1 and 2 fit indexes are on the first and second lines, respectively. For both bifactor models (M4, M4A), the bifactor is orthogonal to the content factors. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; AIC = Akaike's Information Criterion. $n_1 = 607$; $n_2 = 599$

Model 3A: Two Second-Order Factors plus Five Select Correlated Errors Given the good fit of Model 1A, we opted to test the model with two second-order factors (Model 3) plus the correlated errors from Model 1A. In both samples, all measured variables loaded significantly on their respective factors, each first-order factor loaded significantly on its respective second-order factor, and the two second-order factors correlated significantly ($r = .70$). Also, in both samples all estimated correlated errors were significant. (In both samples the disturbance term for factor 5, Personal Alienation, did not have significant variance.) Model fit was appreciably better than Model 3, but because it was more restricted than Models 1 and 2, it did not fit as well.

Model 4: Correlated Content Factors plus a Bifactor This model fit quite well compared to Models 1 and 2, with an appreciably lower AIC. RMSEA was also lower and the CFI and TLI indices were higher. Nearly all the measured variables loaded significantly on their respective content factor and the bifactor. But six of the measured variables did not load significantly on the bifactor and two measured variables did not load significantly on their content factor once the bifactor was added. The variance of all 13 latent variables was significant, but not all covariances among the 12 content factors were significant with the bifactor added. Specifically, with the 12 latent variables allowed to covary, there are 66 possible correlations. In Model 1, the first-order model, all 66 were significant. In Model 4, 39 of the 66 were significant. The largest correlations were for Low Self-Esteem and Personal Alienation ($r = .85$), Interpersonal Insecurity and Interpersonal Alienation ($r = .67$). All others were .50 or below and many were near zero, so overall, adding the bifactor reduced correlations among the 12 content latent variables. It also reduced their variance. For instance, Drive for Thinness's variance decreased 47% and Perfectionism's 41%. Conversely, the variance for Maturity Fears increased with the addition of a bifactor, as did Interpersonal Insecurity, Interpersonal Alienation and Interoceptive Deficits.

Model 4A: Correlated Content Factors plus a Bifactor plus Five Select Correlated Errors Model 4A added five select correlated errors (as described in Model 1A) to Model 4. This was our best fitting model in both samples. All items loaded significantly on their content factor except for three (68, 81, and 86, all of which loaded significantly on the bifactor) and all but five (15, 22, 39, 58, 73) loaded significantly on the bifactor. All factor variances were significant and 39 of 66 estimated factor covariances were significant (as in Model 4). The highest correlations were between Low Self-Esteem and Personal Alienation ($r = .84$) and Interpersonal Alienation and Personal Alienation ($r = .72$). Furthermore, the correlated errors estimated in Model 1A were also significant, ranging from .17 to .57. The pattern was similar for sample 2, with two minor differences: All items loaded significantly on their respective content factors and four (15, 57, 58, 73) did not load significantly on the bifactor.

CFAs for the Combined Samples

The models discussed above were re-tested on the entire sample ($N = 1206$) to provide more stable parameter estimates and fit assessment. Models were chosen either because they were judged to be the best fitting models or because they have precedent in the literature. These models are presented in Table 5. The best fitting model for both samples was Model 4A, which specified correlated content factors plus a bifactor and five select correlated errors. This model had the lowest RMSEA and AIC values and the highest CFI and TLI values as well. In tests of this model on the entire sample, all items loaded on their respective content factor significantly and only four items failed to load significantly on the bifactor. For this best fitting model, standardized loadings of each item on its content factor and the bifactor are presented in columns 5 and 6, respectively, of Table 6.

All latent variable variances were significant, as were the correlated errors (with correlations ranging from .18 to .60). Sixteen of 66 latent variable correlations were non-significant. The two highest factor correlations were between Low Self-Esteem and Personal Alienation (.86) and Personal Alienation and Interpersonal Alienation (.73). It should be noted that these correlations are lower than the observed latent variable correlations without the bifactor present (.92 and .83 respectively). The decision to differentiate the Personal Alienation and Low Self-Esteem scales (and the Interpersonal Insecurity and Interpersonal Alienation scales) in the EDI-3 manual analysis despite the high inter-scale correlations has been made largely on the grounds of clinical utility. Examination of Low Self-Esteem item content indicated that this scale measures negative self-evaluation in contrast to the Personal Alienation scale that assess a more pernicious sense of emotional emptiness. Although there is conceptual and statistical overlap between these scales, there are meaningful clinical differences for patients who score high on one of these scales and low on the other. For example, patients who have high Low Self-Esteem but low Personal Alienation may be more be more amenable to cognitive therapy focusing on negative self-evaluation compared to patients who experience an all-pervasive sense of emotional emptiness reflected by high Personal Alienation. Similarly, examination of item content indicates that patients who have high levels of Interpersonal Insecurity but low Interpersonal Alienation may be more responsive to psychotherapy in general than those whose high Interpersonal Alienation reflects a fundamental lack of trust in relationships.

Table 5 *Goodness-of-Fit Indices for the EDI-3 Best-Fitting Factor Models: Full Sample, Five-Point Scoring*

Model	χ^2	df	CFI	TLI	RMSEA	RMSEA LOW	RMSEA HIGH	AIC
M0: Null	57,152.59	4005						
M1: 12 corr factors	13,602.30	3849	.816	.809	.046	.045	.047	5904.30
M1A: 12 corr factors, select corr errors	12,340.21	3844	.840	.833	.043	.042	.044	4652.21
M2: 12 corr factors, Inconsistency Scale corr errors	13,126.70	3897	.826	.822	.044	.043	.045	5332.70
M3: Two 2nd order factors	14,392.05	3902	.803	.797	.047	.046	.048	6588.05
M3A: Two 2nd order factors, select corr errors	13,126.70	3897	.826	.822	.044	.043	.045	5332.70
M4: Bifactor, corr content factors	11,206.94	3759	.860	.851	.041	.040	.041	3688.94
M4A: Bifactor, corr content factors, select corr errors	9921.80	3754	.884	.846	.037	.036	.038	2413.80

Note: For both bifactor models, (M4, M4A), the bifactor is orthogonal to the content factors. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; AIC = Akaike's Information Criterion. $N = 1206$

IRT Analyses

Item principal components analyses showed that the three risk scales (Drive for Thinness, Bulimia, Body Dissatisfaction) and four of the nine psychological scales (Low Self-Esteem, Personal Alienation, Interpersonal Alienation, Interoceptive Deficits) had first-to-second eigenvalue ratios $\geq$3.0, large enough to conclude that the IRT unidimensionality requirement was not violated (Gomez, 2008). For the five scales that did not reach that threshold (Interpersonal Insecurity, Emotional Dysregulation, Perfectionism, Asceticism, Maturity Fears), the Rasch results (summarized below) point to possible sources of misfit.

Infit/Outfit (Table 6). Of the 90 scored items, eight (1, 11, 19, 40, 47, 53, 72, 81) had infit or outfit mean squares $\geq$1.5, but in only two instances did those values exceed 2.0. And the large mean squares for those items—indicating a lack of coherence with other scale items—are not surprising; item 53 is the only item on the Bulimia scale that asks specifically about thoughts of vomiting to lose weight and item 47 is the only item on Body Dissatisfaction that refers to "feeling bloated" after eating a meal, rather than to dissatisfaction with the size or shape of one's stomach, thighs, etc. Unsurprising as well were the large mean squares for items 72 and 81 on the Emotional Dysregulation, the only items about tendencies to abuse drugs and alcohol, respectively. Consistent with these statistics, this item pair had the largest residual correlation in the CFAs. Finally, as the third column of Table 6 shows, the eight items with mean squares >1.5 had the lowest corrected item-total correlations on their respective scales, but all values were nevertheless $\geq$ .30.

Response Category Thresholds and Ordering With only minor exceptions, there was a reasonable separation of the category response thresholds for the 90 items and there were no items where adjacent response categories were reversed or "disordered."

Person and Item Separation (Table 6). Only two of 12 scales had acceptable person reliability and separation values, whereas all 12 scales exceeded the recommended thresholds for item reliability and separation. The EDI-3 scales are relatively brief (modal scale length = 7 items) and several scales had poor person-item alignment, factors that, as noted earlier, degrade person reliability. The excellent item reliability and separation values were expected, given the large sample size.

Estimated Item Discrimination (Table 6, final column). Only five of the 90 scale-scored items—and no more than one from any individual scale—failed to meet the .50 discrimination threshold and four of the five items that did not reach the threshold (items 1, 40, 47, and 53) were among the eight misfitting items.

Discussion

The results of this study, the first comprehensive investigation of the EDI-3 at the item, scale, and composite levels for an American clinical sample, lead to a positive verdict on its psychometric merits. Confirmatory factor analyses provided support for the 12 scales and a bifactor model defined primarily by the three risk scales, and Rasch analyses identified few problematic items. A detailed summary of our major findings and their implications follows.

Structural Analyses

Several CFA models were tested, based on previous literature and the anticipated factor structure of the EDI-3. In addition to

Table 6 *EDI-3 Item Statistics (N = 1206)*

Item	M	SD	Item-total r	Factor Loading		Rasch Statistics			
				Content	Bifactor	Measure	Infit	Outfit	D
DT (PR = .69, PS = 1.49; IR = .99, IS = 10.30)									
1 (r)	2.71	1.32	.52	.43	.33	0.30	1.35	**1.63**	**0.26**
7	2.73	1.46	.74	.59	.54	0.27	0.81	0.80	1.16
11	3.25	1.32	.59	.42	.48	−0.47	**1.51**	**1.50**	**0.94**
16	3.30	1.22	.78	.65	.50	−0.57	0.80	0.62	1.23
25	2.51	1.46	.60	.46	.45	0.53	1.12	1.14	0.70
32	2.88	1.46	.77	.67	.53	0.09	0.81	0.72	1.31
49	3.05	1.37	.75	.63	.50	−0.15	0.88	0.78	1.21
B (PR = .81, PS = 2.05; IR = .99, IS = 11.65)									
4	1.29	1.47	.80	.71	.44	0.28	0.77	0.88	1.17
5	1.14	1.41	.82	.80	.38	0.49	0.64	0.65	1.26
28	1.37	1.53	.81	.74	.43	0.18	0.79	0.78	1.21
38	1.53	1.56	.83	.58	.35	−0.03	0.73	0.74	1.25
46	1.22	1.44	.79	.75	.37	0.39	0.81	0.78	1.16
53	2.20	1.68	.53	.31	.52	−0.86	**1.90**	**2.43**	**0.29**
61	1.60	1.49	.70	.54	.49	−0.12	1.15	1.32	0.70
64	1.79	1.64	.70	.49	.55	−0.35	1.28	1.25	0.85
BD (PR = .77, PS = 1.82; IR = .99, IS = 11.37)									
2	2.73	1.47	.70	.53	.50	0.05	0.97	1.06	1.01
9	2.69	1.59	.77	.70	.47	0.10	0.88	0.80	1.38
12 (r)	3.07	1.18	.67	.55	.36	−0.35	0.81	1.04	1.01
19 (r)	3.05	1.19	.59	.52	.29	−0.33	1.02	**1.55**	0.83
31 (r)	2.91	1.33	.58	.56	.26	−0.16	1.18	1.38	0.81
45	2.39	1.66	.80	.71	.47	0.41	0.75	0.68	1.45
47	2.73	1.40	.43	.26	.46	0.05	**1.65**	**2.46**	**0.00**
55 (r)	3.12	1.19	.73	.71	.32	−0.42	0.71	0.71	1.24
59	1.97	1.69	.68	.64	.37	0.85	1.07	1.00	1.06
62 (r)	2.95	1.25	.72	.72	.28	−0.20	0.74	0.83	1.13
LSE (PR = .78, PS = 1.89; IR = .99, IS = 12.85)									
10	2.10	1.42	.69	.48	.60	0.22	0.96	0.92	1.11
27	2.26	1.40	.70	.51	.58	0.00	0.92	0.90	1.15
37 (r)	2.75	1.16	.61	.58	.35	−0.70	1.01	1.02	1.01
41	2.59	1.37	.71	.55	.55	−0.46	0.96	0.88	1.16
42 (r)	1.84	1.28	.51	.45	.34	0.56	1.28	1.35	0.56
50 (r)	1.98	1.26	.64	.66	.32	0.39	0.93	0.98	1.04
PA (PR = .76, PS = 1.78; IR = .99, IS = 9.42)									
18	1.95	1.38	.67	.46	.56	0.17	0.85	0.83	1.30
20 (r)	2.49	1.15	.48	.38	.37	−0.44	1.02	1.16	0.85
24	1.87	1.42	.58	.42	.51	0.26	1.08	1.07	0.99
56	1.97	1.39	.63	.42	.57	0.15	0.93	0.92	1.16
80 (r)	1.88	1.19	.42	.39	.28	0.25	1.14	1.28	0.58
84	2.01	1.45	.62	.35	.61	0.10	1.04	1.02	1.08
91 (r)	2.52	1.24	.55	.54	.36	−0.48	0.97	1.04	1.02
II (PR = .76, PS = 1.77; IR = .98, IS = 7.54)									
15 (r)	1.85	1.23	.56	.74	.05	−0.11	0.90	0.88	1.19
23 (r)	1.36	1.17	.62	.65	.12	0.46	0.77	0.76	1.29
34	2.02	1.31	.52	.57	.42	−0.30	1.12	1.13	0.68
57 (r)	1.72	1.19	.58	.76	.05	0.05	0.82	0.81	1.27
69 (r)	1.93	1.30	.57	.47	.21	−0.19	0.98	0.94	1.11
73 (r)	1.57	1.35	.51	.46	−.01	0.22	1.18	1.15	0.89
87	1.88	1.32	.44	.35	.42	−0.13	1.27	1.32	0.52
IA (PR = .72, PS = 1.59; IR = 1.00, IS = 19.55)									
17 (r)	1.84	1.18	.58	.65	.50	−0.28	0.72	0.73	1.32
30 (r)	1.26	1.21	.47	.60	.18	0.34	1.00	1.00	1.05
54	1.79	1.40	.57	.44	.53	−0.23	0.98	0.97	1.04
65	1.57	1.36	.52	.34	.50	0.00	1.04	1.02	0.90
74	0.97	1.20	.39	.24	.40	0.70	1.25	1.19	0.80
76 (r)	2.69	1.10	.39	.43	.23	−1.25	1.05	1.12	0.88
89 (r)	0.95	1.17	.45	.50	.24	0.72	1.13	1.04	1.03
ID (PR = .83, PS = 2.19; IR = .99, IS = 8.35)									
8	2.12	1.50	.62	.28	.62	−0.17	1.18	1.11	0.96

Table 6 (continued)

Item	M	SD	Item-total r	Factor Loading		Rasch Statistics			
				Content	Bifactor	Measure	Infit	Outfit	D
21	1.86	1.33	.73	.66	.49	0.11	0.71	0.70	1.37
26 (r)	1.94	1.18	.57	.67	.28	0.03	0.92	1.17	0.66
33	2.36	1.36	.62	.38	.55	−0.43	1.02	1.03	0.98
40	2.09	1.39	.48	.20	.49	−0.14	1.39	**1.51**	**0.42**
44	2.10	1.46	.65	.29	.65	−0.15	1.03	1.02	1.08
51	1.77	1.38	.72	.58	.52	0.21	0.79	0.76	1.35
60	1.99	1.36	.76	.58	.58	−0.03	0.66	0.64	1.47
77	1.45	1.44	.52	.26	.51	0.57	1.40	1.37	0.67
ED (PR = .65, PS = 1.38; IR = 1.00, IS = 15.43)									
67	1.49	1.40	.56	.42	.54	−0.42	0.84	0.80	1.08
70	1.21	1.24	.47	.42	.34	−0.17	0.93	1.02	0.79
72	0.41	1.04	.35	.14	.22	0.90	**1.68**	1.41	1.07
79	0.96	1.25	.54	.61	.30	0.08	0.94	0.86	1.07
81	0.51	1.12	.32	.10	.23	0.70	**1.67**	1.42	1.00
83	1.64	1.42	.58	.62	.40	−0.55	0.81	0.80	1.09
85	1.96	1.40	.57	.42	.61	−0.82	0.77	0.78	0.99
90	0.79	1.22	.35	.18	.36	0.29	1.38	1.45	0.88
P (PR = ..71, PS = 1.58; IR = 1.00, IS = 17.10)									
13	1.37	1.41	.57	.42	.27	0.89	0.96	0.98	1.04
29	2.87	1.34	.52	.48	.26	−0.70	1.16	1.12	0.95
36	2.62	1.38	.62	.70	.41	−0.42	0.86	0.87	1.18
Item	M	SD	Item-total r	Factor Loading Content	Bifactor	Rasch Statistics Measure	Infit	Outfit	D
43	1.78	1.43	.54	.37	.22	0.45	1.02	1.03	0.89
52	2.27	1.43	.62	.66	.46	−0.05	0.87	0.87	1.13
63	2.39	1.44	.50	.54	.21	−0.17	1.18	1.32	0.76
A (PR = .72, PS = 1.62; IR = 1.00, IS = 22.28)									
66	2.34	1.46	.59	.31	.62	−0.45	0.82	0.84	1.08
68	3.15	1.26	.47	.16	.53	−1.29	1.07	1.20	1.00
75	1.05	1.34	.44	.33	.38	0.67	1.09	1.19	0.70
78	1.78	1.55	.55	.45	.46	0.02	0.96	0.93	1.15
82	1.20	1.32	.36	.46	.25	0.52	1.18	1.42	0.70
86	1.90	1.56	.53	.12	.64	−0.08	1.01	0.98	1.12
88	0.95	1.17	.48	.42	.39	0.61	0.90	0.97	0.95
MF (PR = .76, PS = 1.77; IR = .99, IS = 8.79)									
3	1.43	1.45	.66	.67	.32	0.03	1.02	0.95	1.20
6	0.99	1.29	.62	.59	.35	0.57	1.07	1.01	1.12
14	1.35	1.36	.66	.68	.29	0.12	0.91	0.88	1.18
22 (r)	1.48	1.32	.55	.69	−.10	−0.03	1.05	1.04	1.05
35	1.56	1.30	.38	.29	.53	−0.12	1.37	1.45	**0.39**
39 (r)	1.64	1.26	.54	.68	−.07	−0.20	0.95	0.97	1.06
48	1.37	1.24	.58	.57	.32	0.10	0.91	0.95	0.95
58 (r)	1.90	1.09	.49	.58	−.02	−0.47	0.83	0.92	0.94

Note. For all items, higher scores reflect greater distress. Item numbers followed by (r) are reverse-scored. Bolded mean squares indicate significant misfit (Linacre, 2016). D = Estimated discrimination. DT = Drive for Thinness; B = Bulimia; BD = Body Dissatisfaction; LSE = Low Self-Esteem; PA = Personal Alienation; II = Interpersonal Insecurity; IA = Interpersonal Alienation; ID = Interoceptive Deficits; ED = Emotional Dysregulation; P = Perfectionism; A = Asceticism; MF = Maturity Fears. PR = Person reliability; PS = Person separation; IR = Item reliability; IS = Item separation

previous work, we chose to test a bifactor model. The best fitting model was a bifactor model with all 12 content factors allowed to correlate and five select correlated errors. The RMSEA as well as the upper bound (90% CI) of the RMSEA were below .04. The incremental fit indices (CFI and TLI) were below the commonly accepted cut-offs. A more ideal outcome with respect to validating this measure would have been for these indices to exceed .90 or .93. However, a common criticism of incremental fit indices is that they are reliant on the badness of fit of the Null Model (Kline, 2016). While there is a general positive manifold to these data, many of the item covariances are relatively low.

A second issue concerns the structure of the data vis a vis the bifactor. While the bifactor model provides very good fit to the data and is thought to represent general distress in this clinical population, it is not a pragmatic model from the

practitioner's point of view. We propose that the bifactor model could be best used in research utilizing this measure, especially when SEM methods are used for analyses. The bifactor could be used as a latent variable in the analysis along with the content factors, separating general distress from the constructs measured by the content factors. We further recommend that researchers using the EDI-3 in research studies utilizing SEM consider using the Exploratory SEM method proposed by Asparouhov and Muthén (2009).

Finally, we view the results as generally supporting the purported primary scale structure of the EDI-3. Without considering correlated errors, or the bifactor solution, the best fitting model was the 12 correlated factors model.

Item Analyses

The three risk scales and four of the nine psychological scales met the IRT criterion for unidimensionality. Five scales did not. However, given the small number of misfitting items (eight), adequate item-total correlations, strong estimated discrimination values for all but five items, and the generally large item loadings on the bifactor and/or group factors, we concluded that violations of the Rasch unidimensionality requirement were substantively inconsequential.

Of the eight items that did not fit the Rasch model, six were underfit, the result of being the only reverse-worded item on a scale (1) or the only item referencing a specific feeling or eating disorder-related behavior (11, 19, 40, 47, 53). Two Emotional Dysregulation scale items (72, 81) were overfit, the result of their being the only two substance abuse items on the scale. This also explains why these two items, which form the "substance abuse risk" item cluster described by Garner (2004), had the largest correlated residuals in the CFAs. Although these items do not fit the Rasch model, they may have clinical utility in identifying potential substance abuse problems.

Eight items are thus potential candidates for deletion or replacement in a revision of the EDI-3. However, in each instance the degree of misfit was not severe, all eight items had corrected item-total correlations $\geq$ .30, and four of the eight had adequate estimated discrimination values. Furthermore, an item that is a scale's sole indicator of a symptom (e.g., vomiting), despite its psychometric shortcomings, may nevertheless have value as a "critical item." For these reasons, and because the sources of misfit were identifiable and correctable, minor changes in item phrasing should be considered before deletion or replacement.

Clinical Implications

The EDI-3 and earlier versions of the test were formulated using an approach to construct validation relying on both rational and empirical methods of scale development. Scales were originally generated by experienced clinicians based upon constructs derived from the clinical literature and scales were retained if they demonstrated stability and empirical adequacy. At the scale level, the evidence reported here largely confirms the psychometric properties reported in the manual (Garner, 2004). Scale mean scores and reliability coefficients for the current sample were comparable to those reported in the manual for the clinical normative samples (Garner, 2004). It is important to note that the results of the current study speak to only one aspect of construct validity but add to the body of evidence for other dimensions such as convergent, discriminant, predictive and concurrent validity described in the EDI-3 manual (Garner, 2004)..

Similarly, the CFA results are generally supportive of the 12 content scales described by Garner (2004) in the manual as first-order factors with the best fitting model specifying a general psychological distress bifactor on which most of the items load, in addition to loading on their respective first-order scales. However, a model with second-order factors corresponding to the risk and psychological composites has heuristic merit since it is consistent with the EDI-3 conceptual framework (Garner, 2004) as well as the findings reported by Clausen et al., 2011. The configuration of scales around these two broad domains of eating disorder risk and psychological features may have the greatest interpretive value for clinicians. The first composite relates to symptoms specific to those with clinical or subclinical eating disorders, whereas the second composite assesses psychological features useful in case conceptualization, treatment planning, and assessment of progress. This type of information is particularly relevant in individual cases because it is recognized that patients vary remarkably within diagnostic groups on the psychological dimensions assessed by the EDI-3.

We did not assess scale and composite differences among eating disorder diagnostic groups because such differences may have limited interpretive value due to the considerable within group variance on the EDI-3 scales. Diagnostic groups are at best relatively crude differentiations between patients based on current symptoms and weight. It is well-established that eating disorder patients move between diagnostic groups at different points in time. This contrasts with the relative stability of the traits measured by the EDI-3 (Joiner, Heatherton, & Keel, 1997; Baxter, Heatherton, & Joiner, 2007; Rizvi, Stice, & Agras, 1999). Although there are significant improvements in the EDI-3 with treatment, pre-to-post-treatment change scores are not significantly correlated with amount or rate of weight gain in anorexia nervosa (Garner, Desmond, Desai, & Lockert, 2016). However, there are moderate correlations between pretreatment and post treatment EDI-3 scores indicating relative trait stability in contrast to shifts in symptoms over time (Garner, in preparation).

One limitation of the current study relates to the eating disorder sample, which consisted primarily of Caucasian female patients admitted to a partial hospitalization or residential in Ohio with a diagnosis of either anorexia nervosa or bulimia nervosa. Therefore, the findings may not be generalizable to other patient groups such as males, non-whites, or those with other eating disorder diagnoses such as BED, OSFED, or milder cases. Nevertheless, while the pertinence of the content domains assessed by the EDI-3 may vary across populations, leading to different norms, the item clusters themselves (the primary focus of the current study) show robust factorial association and internal consistency. Another limitation of the current study was the need to impute data for EDI-3 questions that were left blank. There are limitations for all data imputation methods, and this extends to Maximum Likelihood methods such as the one used here (EM). First it is assumed that one has a large data set to work from when computing EM estimates to ensure those estimates are approximately unbiased and it is also assumed that data are Missing at Random (Shafer and Graham, 2002). Further, these assume an underlying parametric model that gives rise to the data and it is not always clear whether this model holds for the missing values, though a saturated model for EM is available as well and was used here. Finally, some researchers believe that ML methods perform better under conditions of multivariate normality (see Gold & Bentler, 2000, for further discussion).

Conclusions

This study is the first comprehensive psychometric investigation of the EDI-3 for an American patient sample. Consistent with Clausen et al.'s (2011) findings for Danish patients and non-patients, we found support for a model specifying first-order factors corresponding to the 12 scales. Clausen et al. then declared their preference for a model with two second-order factors (Garner's [2004] model), arguing that it provided the most parsimonious representation of the data. However, Clausen et al. did not fit bifactor models, nor did they identify any of the "…hugely correlated residuals…" (p. 107) that degraded overall model fit. So a second strength of this study was the specification of additional factor models that included a bifactor and substantively meaningful correlated errors, models that the large patient sample allowed us to cross-validate. Doing so led to our conclusion that a bifactor model with five select correlated errors produced slightly better fit than did the second-order models but, as noted earlier, the latter may have greater utility for practicing clinicians. A third strength of this study is that it was the first to report IRT results for the EDI-3 items. Those analyses identified few statistically problematic items, and for those few items, the problems were generally minor.

Finally, Garner (2004) has always maintained that the EDI-3 is not a diagnostic instrument. "Rather, it is aimed at the *measurement of psychological traits or symptom clusters* relevant to the development and maintenance of eating disorders." (p. 4, author's italics) Garner's disclaimer notwithstanding, the extent to which EDI-3 scores discriminate between those with and without eating disorder diagnoses and the relative contributions of the 12 scales and six composites to such discriminations, is relevant information for researchers and mental health professionals. For their sample of Danish adults, Clausen et al. (2011) ran receiver operating characteristic (ROC) analyses to assess the accuracy of the 12 EDI-3 scales for discriminating normal controls from three DSM-IV (American Psychiatric Association, 1994) diagnostic groups: Anorexia nervosa (AN), Bulimia Nervosa (BN), and Partial AN/BN (for reasons not stated, the EDI-3 composites were not analyzed). Interestingly, the Interoceptive Deficits scale had the highest sensitivity and specificity for classifying AN and Partial AN/BN cases, whereas the Bulimia scale best predicted the BN diagnosis. Our sample included only clinical cases, which precluded ROC analyses. And as noted earlier, there is considerable within-group variability on the EDI-3 scales, which makes doubtful the identification of distinct diagnostic group profiles. Nevertheless, studies that include both clinical cases (based on DSM-V criteria) and matched controls would provide insight into the scales and composites that help differentiate eating disorder patients from non-patients. Moreover, cluster analytic studies may reveal psychological typologies that relate meaningfully to the utility of specific treatment modalities.

References

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

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397–438. https://doi.org/10.1080/10705510903008204.

Brown, T. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.

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

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 180–225. https://doi.org/10.1207/s15327906mbr4102_5.

Clausen, L., Rosenvinge, J. H., Friborg, O., & Rokkedal, K. (2011). Validating the eating disorder Inventory-3 (EDI-3): A comparison between 561 female eating disorders patients and 878 females from the general population. *Journal of Psychopathology and Behavioral Assessment, 33*, 101–110. https://doi.org/10.1007/s10862-010-9207-4.

Cumella, E. J. (2006). Review of the eating disorder Inventory-3. *Journal of Personality Assessment, 87*, 116–117. https://doi.org/10.1207/s15327752jpa8701_11.

Garner, D. M., Olmsted, M. P., & Polivy, J. (1983). Development and validation of a multidimensional eating disorder inventory for anorexia nervosa and bulimia. *International Journal of Eating Disorders, 2*, 15–34. https://doi.org/10.1002/1098-108X(198321)2:2<15::AID-EAT2260020203>3.0.CO;2-6.

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

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

Garner, D. M., Desmond, M., Desai, J., & Lockert, J. (2016). The disconnect between treatment outcome data and reimbursement for the treatment of anorexia nervosa. *International Journal of Physiatry, 2*(006), 10.23937/2572-4215.1510006.

Gleaves, D. H., Pearson, C. A., Ambwani, S., & Morey, L. C. (2014). Measuring eating disorder attitudes and behaviors; a reliability generalization study. *Journal of Eating Disorders, 2*, 1–12. https://doi.org/10.1186/2050-2974-2-6.

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation maximization. *Structural Equation Modeling, 7*, 319–355. https://doi.org/10.1207/S15328007SEM0703_1.

Gomez, R. (2008). Parent rating of the ADHD items of the disruptive behavior rating scale: Analyses of their IRT properties based on the generalized partial credit model. *Personality and Individual Differences, 45*, 181–186. https://doi.org/10.1016/j.paid.2008.04.001.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:Q hypothesis. *Structural Equation Modeling, 10*, 128–141. https://doi.org/10.1207/S15328007SEM1001_6.

Jackson, D. L., Voth, J., & Frey, M. P. (2013). A note on sample size and solution propriety for confirmatory factor analytic models. *Structural Equation Modeling, 20*, 86–97. https://doi.org/10.1080/10705511.2013.742388.

Joiner, T. E., Heatherton, T. F., & Keel, P. K. (1997). Ten-year stability and predictive validity of five bulimia-related indicators. *American Journal of Psychiatry, 154*, 1133–1138. https://doi.org/10.1176/ajp.154.8.1133.

Keel, P. K., Baxter, M. G., Heatherton, T. F., & Joiner, T. E. (2007). A 20-year longitudinal study of body weight, dieting, and eating disorder symptoms. *Journal of Abnormal Psychology, 116*, 422–432. https://doi.org/10.1037/0021-843X.116.2.422.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

Linacre, J. M. (2016). *Winsteps* (version 3.92.1) [software]. Available from http://winsteps.com/winsteps.htm.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667–696. https://doi.org/10.1080/00273171.2012.715555.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 3*, 369–379. https://doi.org/10.1080/10705519609540052.

Rizvi, S. L., Stice, E., & Agras, W. S. (1999). Natural history of disordered eating attitudes over a 6-year period. *International Journal of Eating Disorders, 26*, 406–413. https://doi.org/10.1002/(SICI)1098-108X(199912)26:4<406::AID-EAT6>3.0.CO;2-6.

Shafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177. https://doi.org/10.1037/1082-989X.7.2.147.

Wright, B. D., & Linacre, J. M (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 370.